1-2012

Role of Semantic Web in Health Informatics

Satya S. Sahoo
Wright State University - Main Campus

Guo-Qiang Zhang

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

SCHOOL OF MEDICINE
CASE WESTERN RESERVE UNIVERSITY

Role of Semantic Web in Health Informatics

Tutorial at 2012 ACM SIGHIT International Health Informatics Symposium (IHI 2012), January 28-30, 2012

Satya S. Sahoo, GQ Zhang Amit Sheth

Division of Medical Informatics Kno.e.sis Center

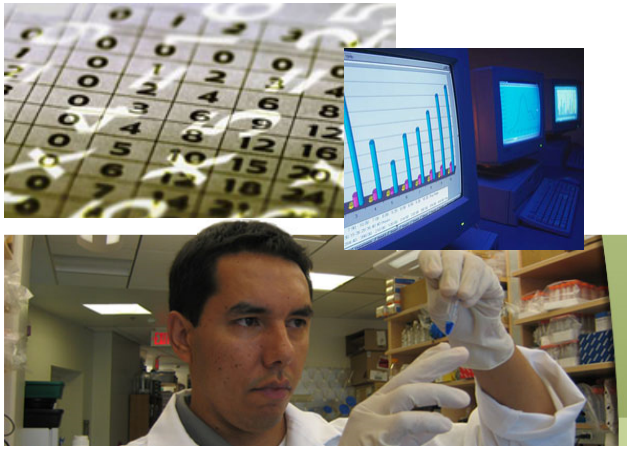
Case Western Reserve University Wright State University

Outline

- Semantic Web
  - Introductory Overview
- Clinical Research
  - Physio-MIMI
- Bench Research and Provenance
  - Semantic Problem Solving Environment for *T.cruzi*
- Clinical Practice
  - Active Semantic Electronic Medical Record

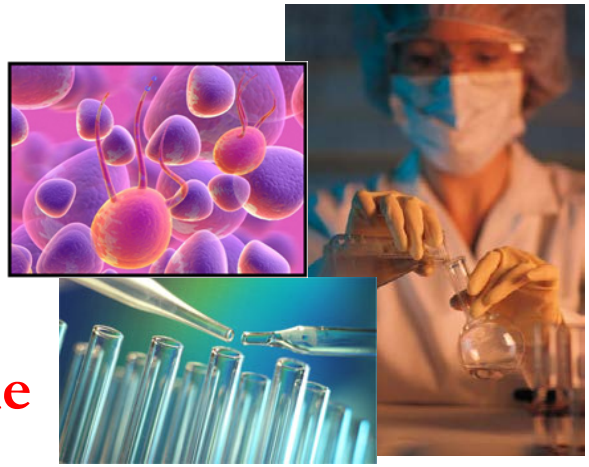
Semantic Web

Landscape of Health Informatics



Clinical Research

- Patient Care
- Personalized Medicine
- Drug Development
- Privacy
- Cost



Bench Research

Clinical Practice

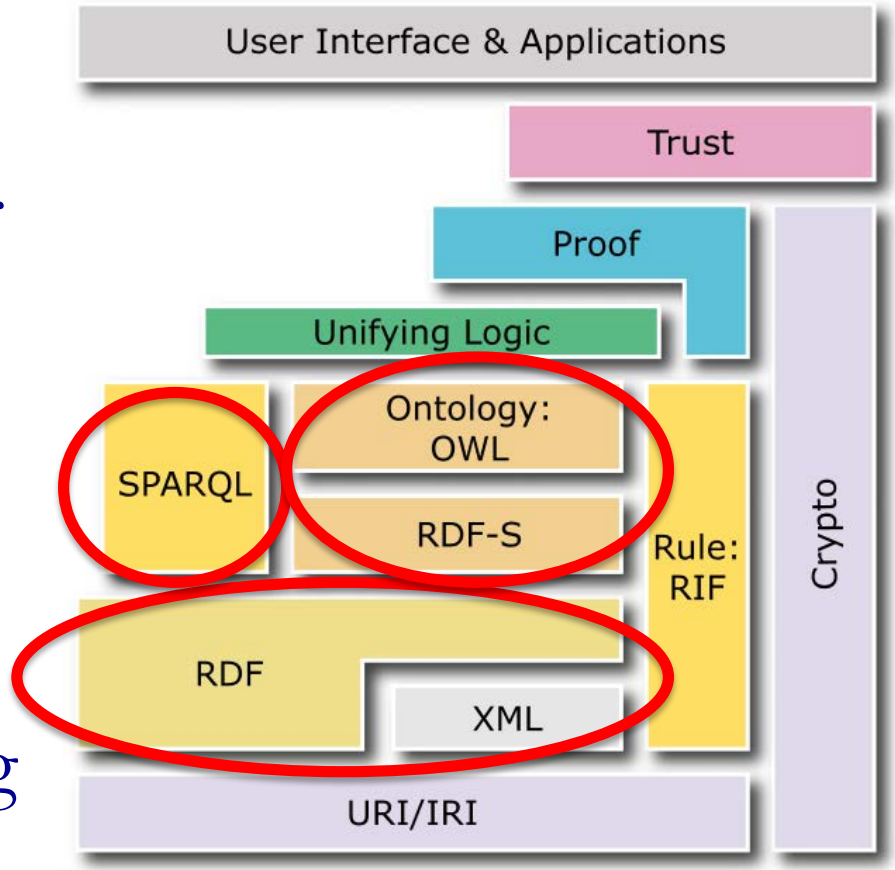
* Images from case.edu

Challenges

- Information Integration: Reconcile heterogeneity
  - Syntactic Heterogeneity: DOB vs. Date of Birth
  - Structural Heterogeneity: Street + Apt + City vs. Address
  - Semantic Heterogeneity: Age vs. Age at time of surgery vs. Age at time of admission
- Humans can (often) accurately interpret, but extremely difficult for machine
  - Role for Metadata/Contextual Information/Semantics

Semantic Web

- Web of Linked Data
- Introduced by Berners Lee et. al as next step for Web of Documents
- Allow "machine understanding" of data,
- Create "common" models of domains using formal language - ontologies



Semantic Web Layer Cake

Layer cake image source: http://www.w3.org

Resource Description Framework

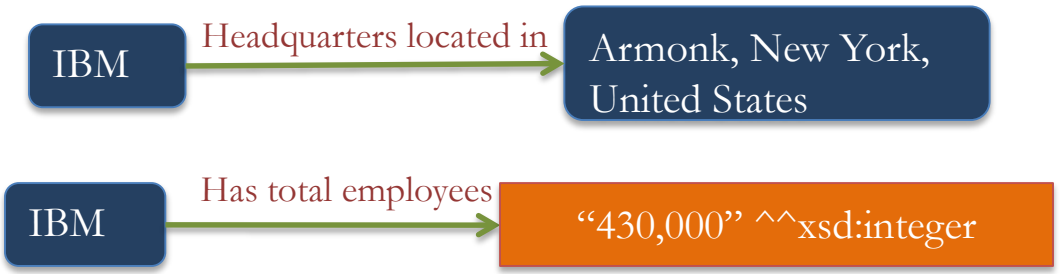
- Resource Description Framework – Recommended by W3C for metadata modeling [RDF]
- A standard common modeling framework – usable by humans and machine understandable

RDF: Triple Structure, IRI, Namespace

- RDF Triple
  - Subject: The resource that the triple is about
  - Predicate: The property of the subject that is described by the triple
  - Object: The value of the property
- Web Addressable Resource: Uniform Resource Locator (URL), Uniform Resource Identifier (URI), Internationalized Resource Identifier (IRI)
- Qualified Namespace: http://www.w3.org/2001/XMLSchema# as xsd:
  - xsd: string instead of http://www.w3.org/2001/XMLSchema#string

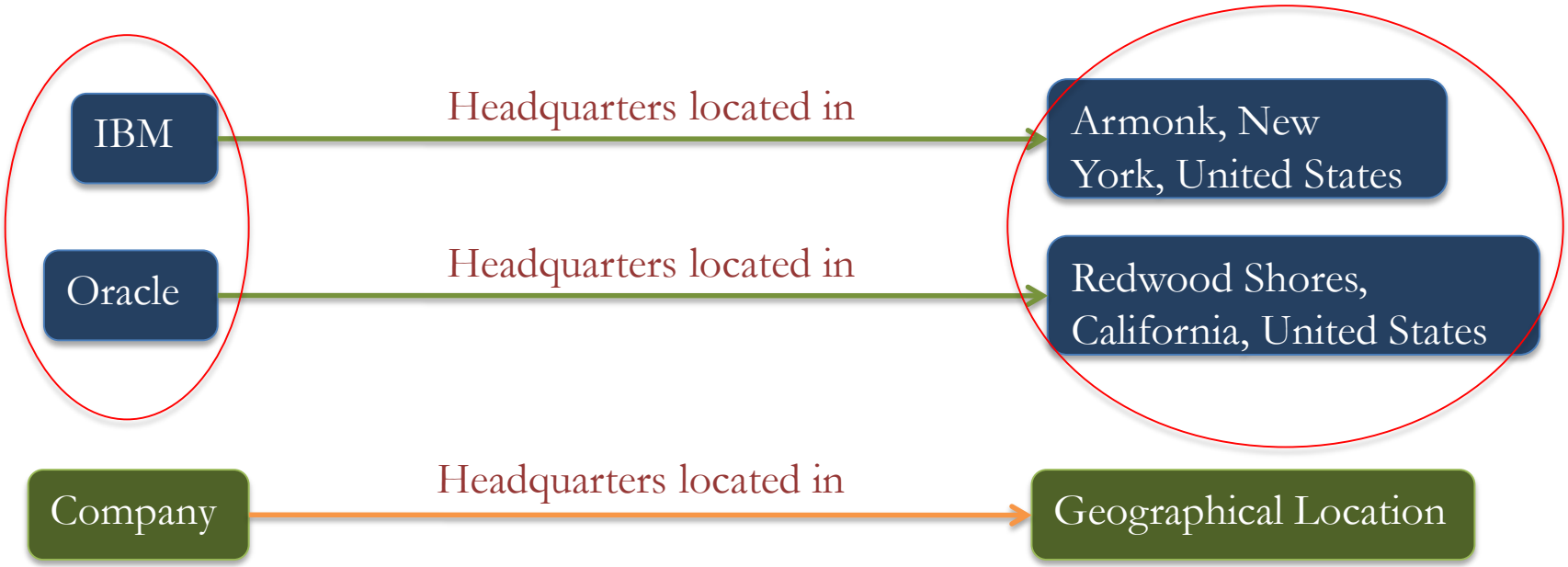
RDF Representation

- Two types of property values in a triple
  - Web resource
  - Typed literal



- The graph model of RDF: node-arc-node is the primary representation model
- Secondary notations: Triple notation
  - companyExample:IBM companyExample:has-Total-Employee "430,000"^^xsd:integer .

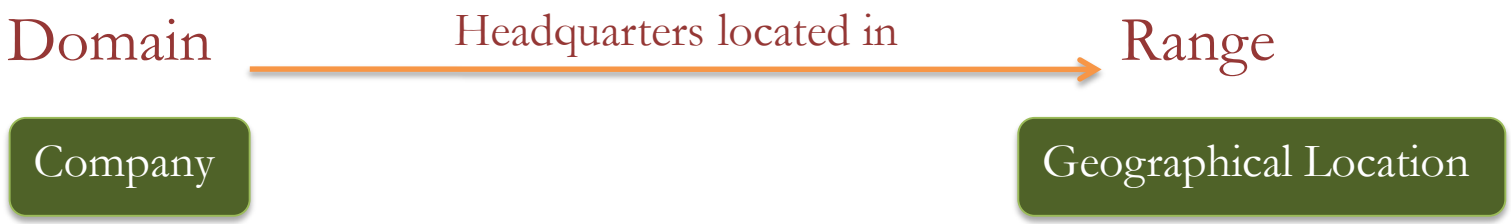
RDF Schema



- RDF Schema: Vocabulary for describing groups of resources [RDFS]

RDF Schema

- Property domain (rdfs:domain) and range (rdfs:range)



- Class Hierarchy/Taxonomy: rdfs:subClassOf

Ontology: A Working Definition

- Ontologies are shared conceptualizations of a domain represented in a formal language*
- Ontologies in health informatics:
  - Common representation model - facilitate interoperability, integration across different projects, and enforce consistent use of terminology
  - Closely reflect domain-specific details (*domain semantics*) essential to answer end user
  - Support reasoning to discover implicit knowledge

* Paraphrased from Gruber, 1993

OWL2 Web Ontology Language

- A language for modeling ontologies [OWL]
- OWL2 is declarative
- An OWL2 ontology (schema) consists of:
  - Entities: Company, Person
  - Axioms: Company employs Person
  - Expressions: A Person Employed by a Company = CompanyEmployee
- Reasoning: Draw a conclusion given certain constraints are satisfied
  - RDF(S) Entailment
  - OWL2 Entailment

OWL2 Constructs

- Class Disjointness: Instance of class A cannot be instance of class B
- Complex Classes: Combining multiple classes with set theory operators:
  - Union: Parent = ObjectUnionOf (:Mother :Father)
  - Logical negation: UnemployedPerson = ObjectIntersectionOf (:EmployedPerson)
  - Intersection: Mother = ObjectIntersectionOf (:Parent :Woman)

OWL2 Constructs

- Property restrictions: defined over property
- Existential Quantification:
  - Parent = ObjectSomeValuesFrom (:hasChild :Person)
  - To capture incomplete knowledge
- Universal Quantification:
  - US President = objectAllValuesFrom (:hasBirthPlace United States)
- Cardinality Restriction

SPARQL: Querying Semantic Web Data

- A SPARQL query pattern composed of triples
- Triples correspond to RDF triple structure, but have variable at:
  - Subject: ?company ex:hasHeadquaterLocation ex:NewYork.
  - Predicate: ex:IBM ?whatislocatedin ex:NewYork.
  - Object: ex:IBM ex:hasHeadquaterLocation ?location.
- Result of SPARQL query is list of values – values can replace variable in query pattern

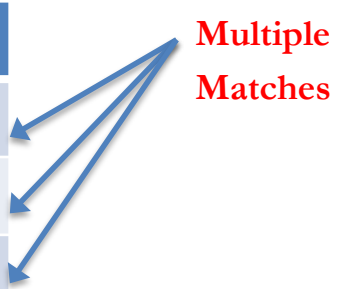
SPARQL: Query Patterns

- An example query pattern

```
PREFIX ex:<http://www.eecs600.case.edu/>
SELECT ?company ?location WHERE
{?company ex:hasHeadquaterLocation ?location.}
```

- Query Result

company	location
IBM	NewYork
Oracle	RedwoodCity
MicorosoftCorporation	Bellevue



SPARQL: Query Forms

- SELECT: Returns the values bound to the variables
- CONSTRUCT: Returns an RDF graph
- DESCRIBE: Returns a description (RDF graph) of a resource (e.g. IBM)
  - The contents of RDF graph is determined by SPARQL query processor
- ASK: Returns a Boolean
  - True
  - False

Semantic Web + Clinical Research Informatics =

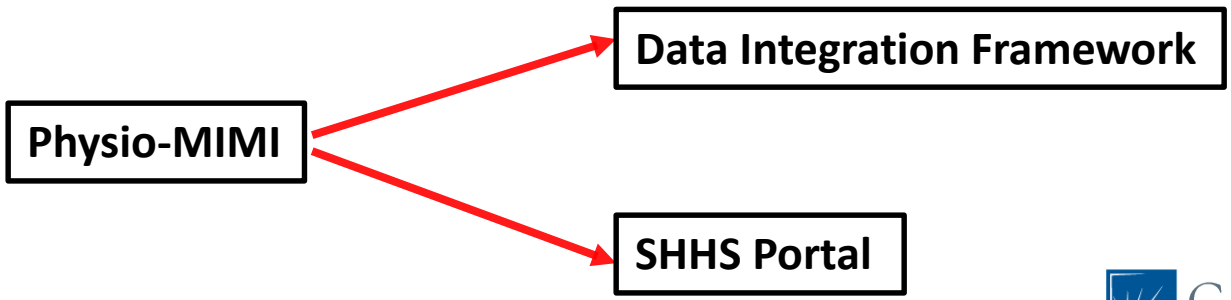
Physio-MIMI

Physio-MIMI Overview

- Physio-MIMI: Multi-Modality, Multi-Resource Environment for Physiological and Clinical Research
- NCRR-funded, multi-CTSA-site project (RFP 08-001) for providing informatics tools to clinical investigators and clinical research teams at and across CTSA institutions to enhance the collection, management and sharing of data
- Collaboration among Case Western, U Michigan, Marshfield Clinic and U Wisconsin Madison
- Use Sleep Medicine as an exemplar, but also generalizable
- Two year duration: Dec 2008 – Dec 2010

Features of Physio-MIMI

- Federated data integration environment
  - Linking existing data resources without a centralized data repository
- Query interface directly usable by clinical researchers
  - Minimize the role of the data-access middleman
- Secure and policy-compliant data access
  - Fine-grained access control, dual SSL, auditing
- Tools for curating PSGs

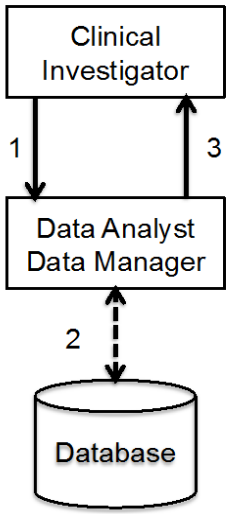


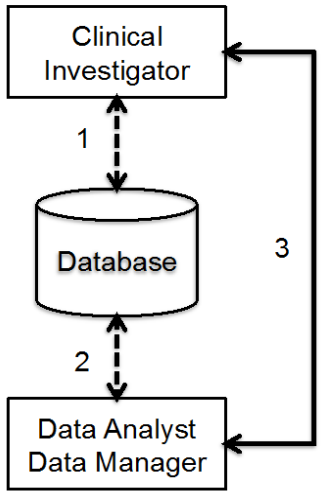
Data Access, Secondary Use

Measure not by the size of the database, but the number of secondary studies it supported

Query Interface – driven by access

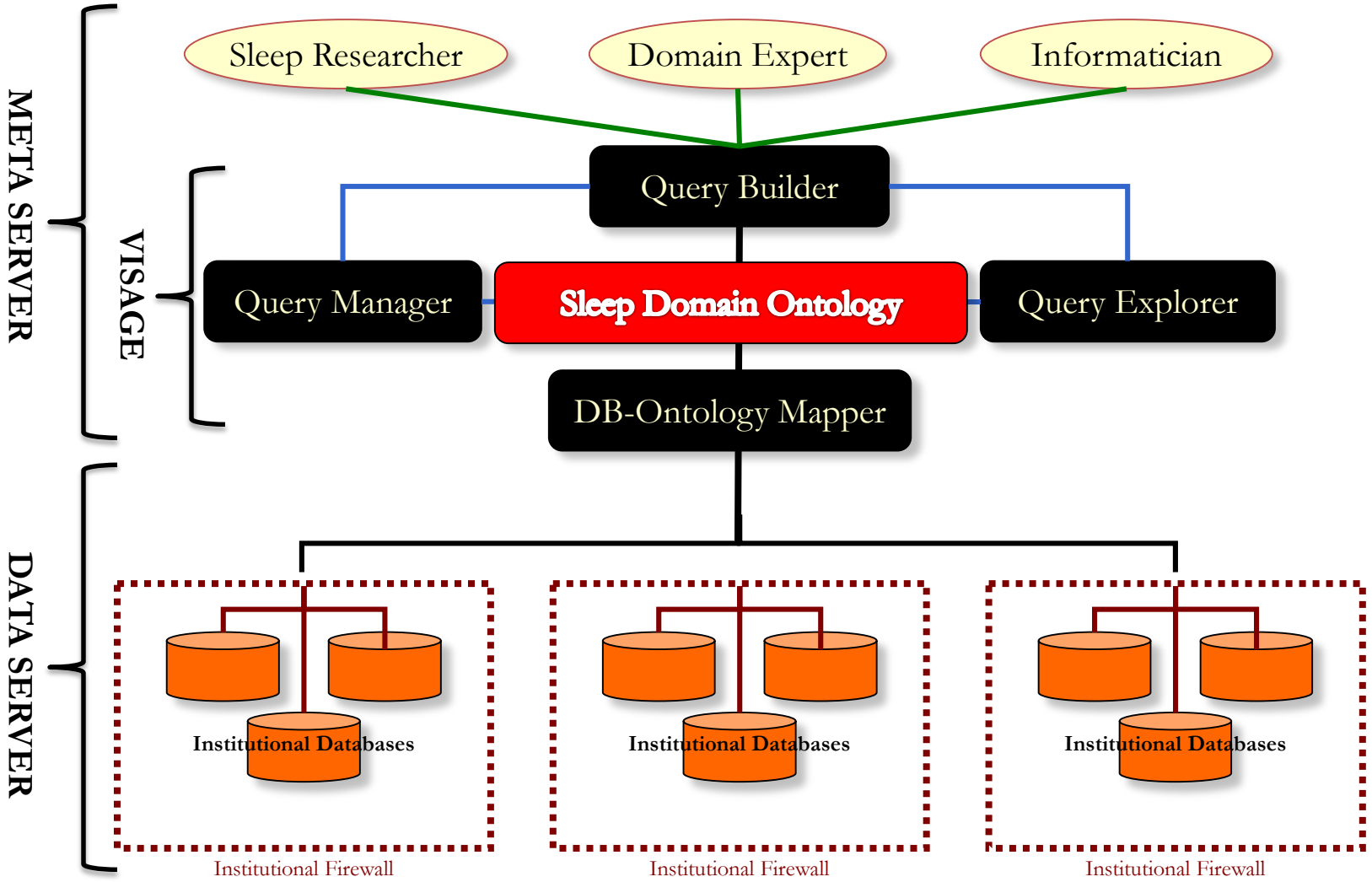
- Visual Aggregator and Explorer (VISAGE)
- Federated, Web-based
- Driven by Domain Ontology (SDO)
- PhysioMap to connect autonomous data sources



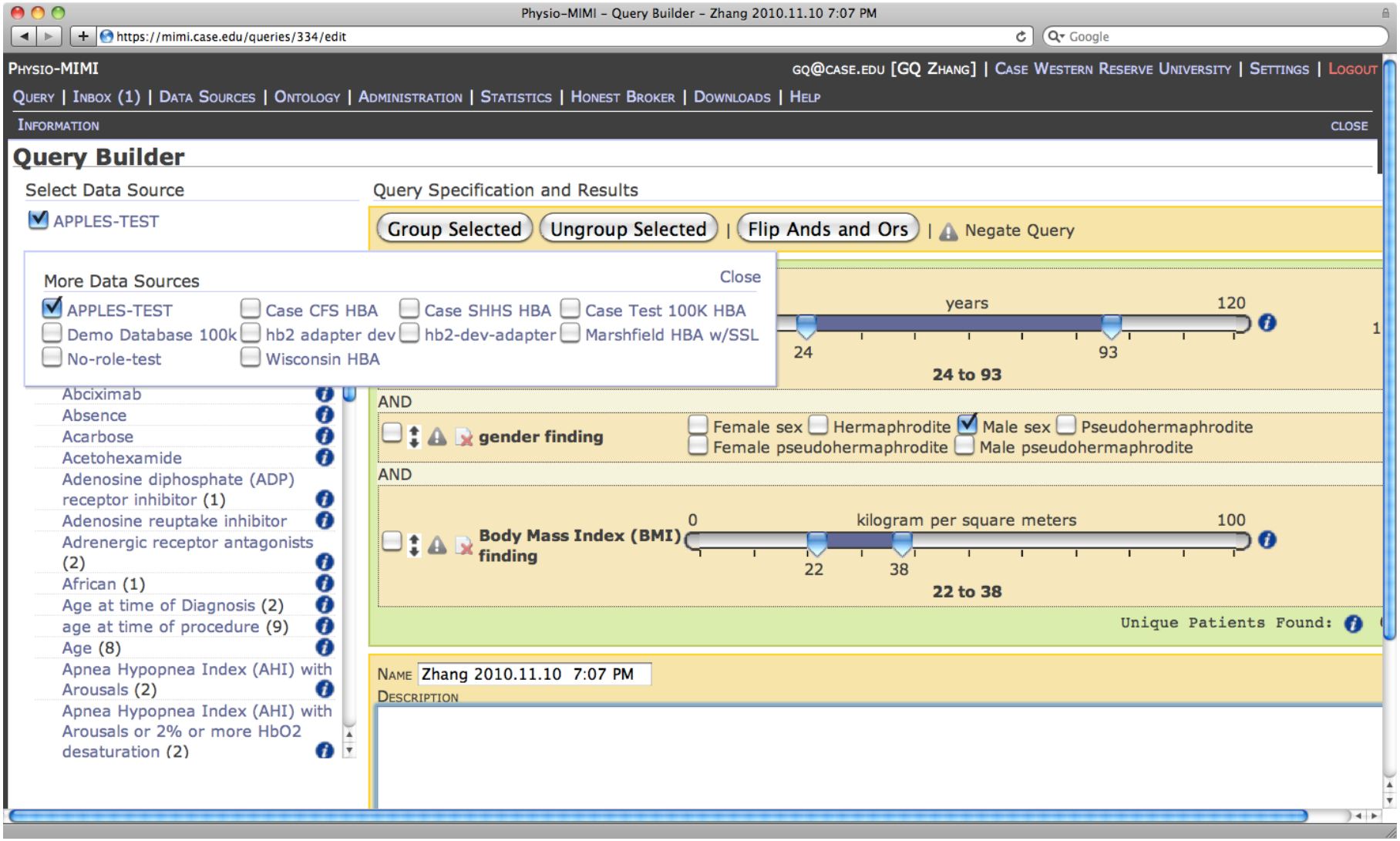


- GQ Zhang et al.
  VISAGE: A Query Interface for Clinical Research, Proceedings of the 2010 AMIA Clinical Research Informatics Summit, San Francisco, March 12-13, pp. 76-80, 2010

Physio-MIMI Components



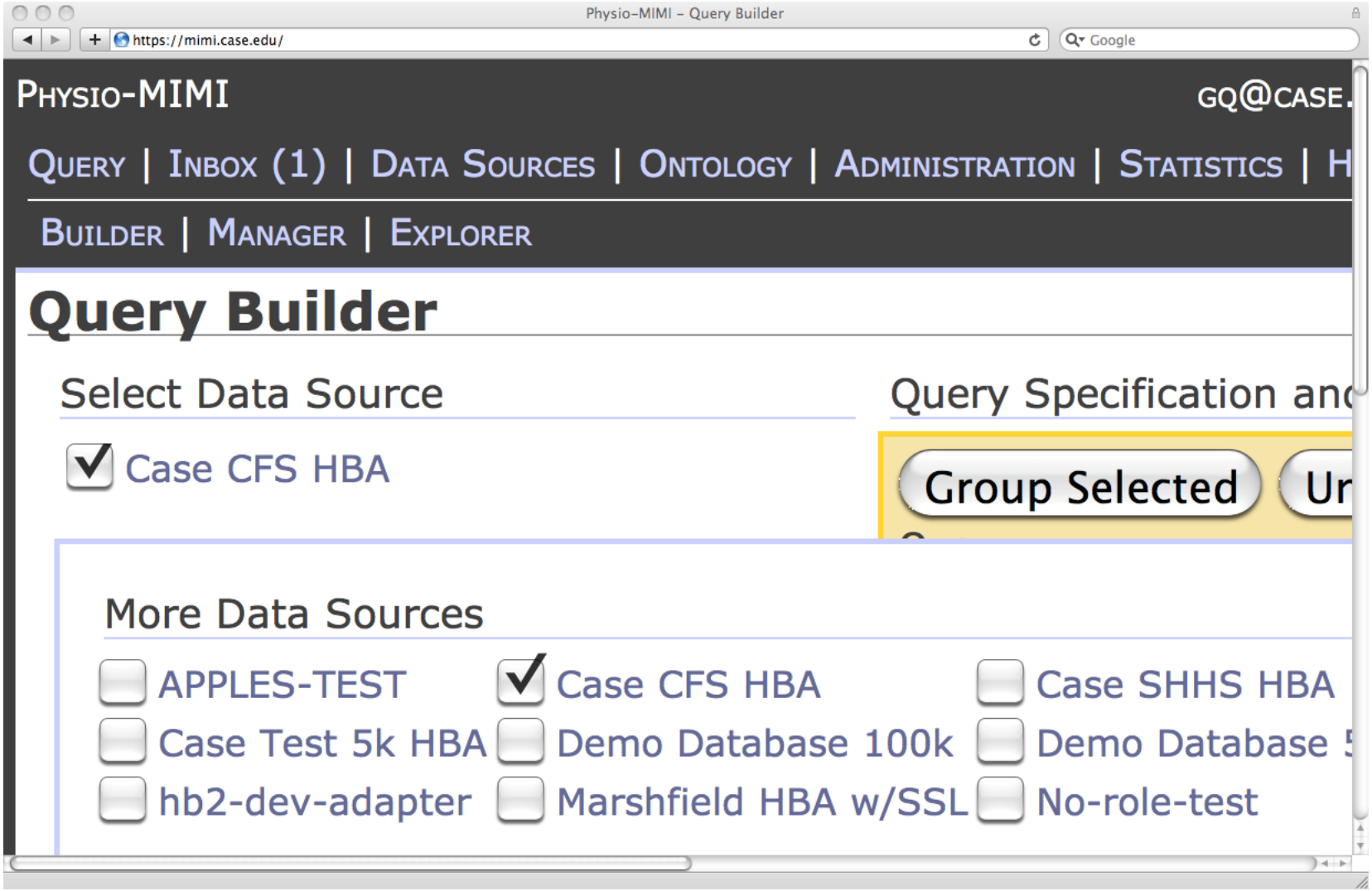
VISAGE screenshot



kno.e.sis

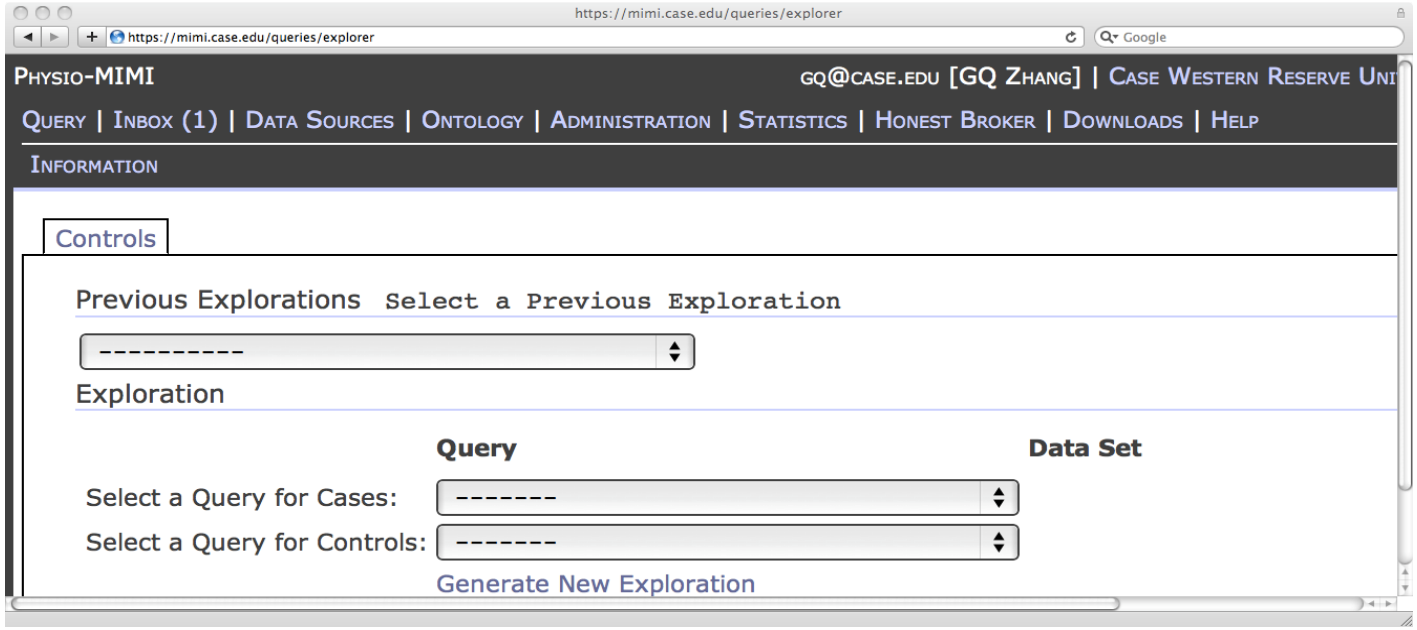
CASE WESTERN RESERVE UNIVERSITY
SCHOOL OF MEDICINE

Components of VISAGE



Case Control Study Design

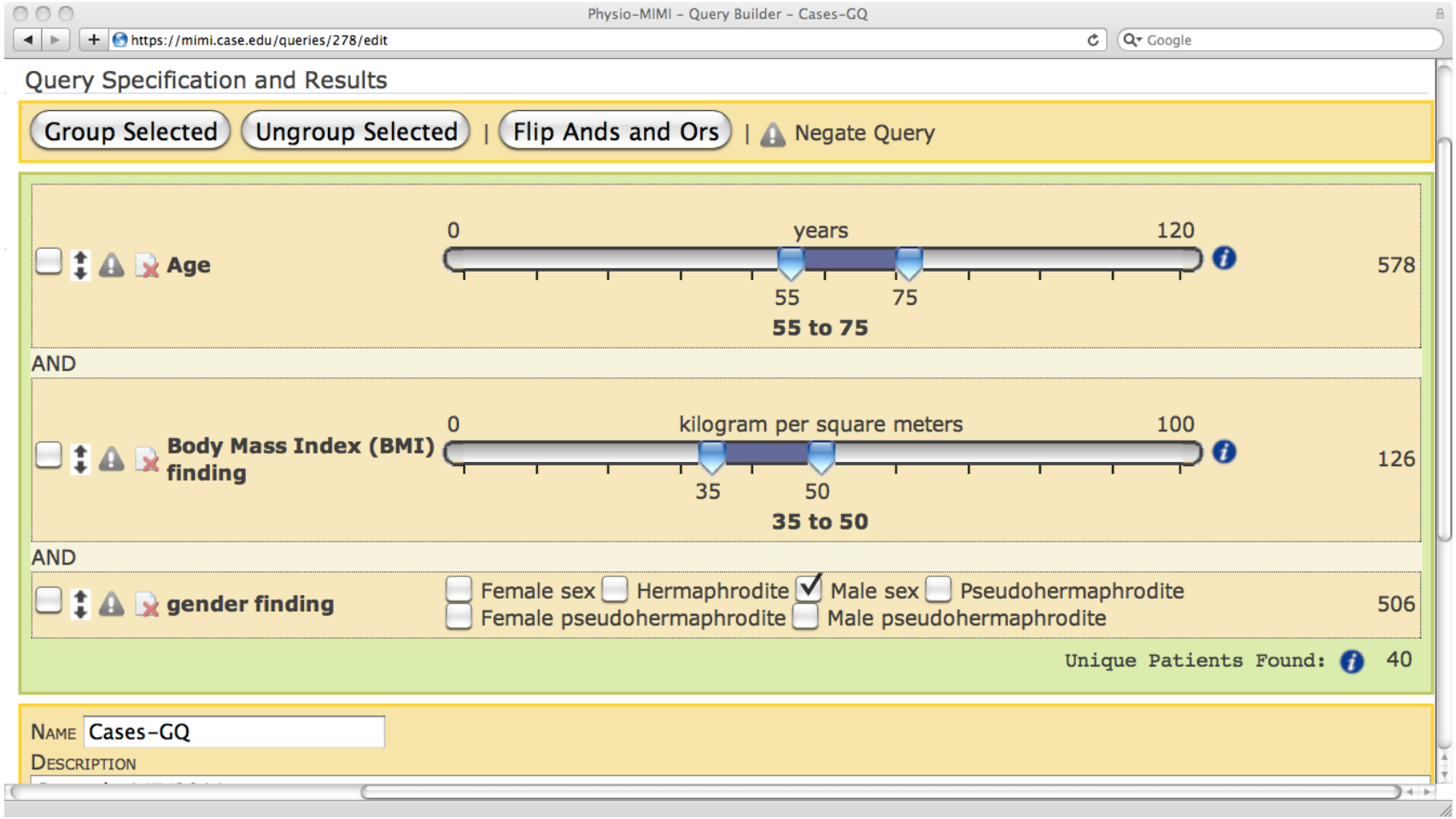
- Case-control is a common study design
- Used for epidemiological studies involving two cohorts, one representing the **cases**
and the second representing the **controls**
- Adjusting matching ratio to improve statistical power



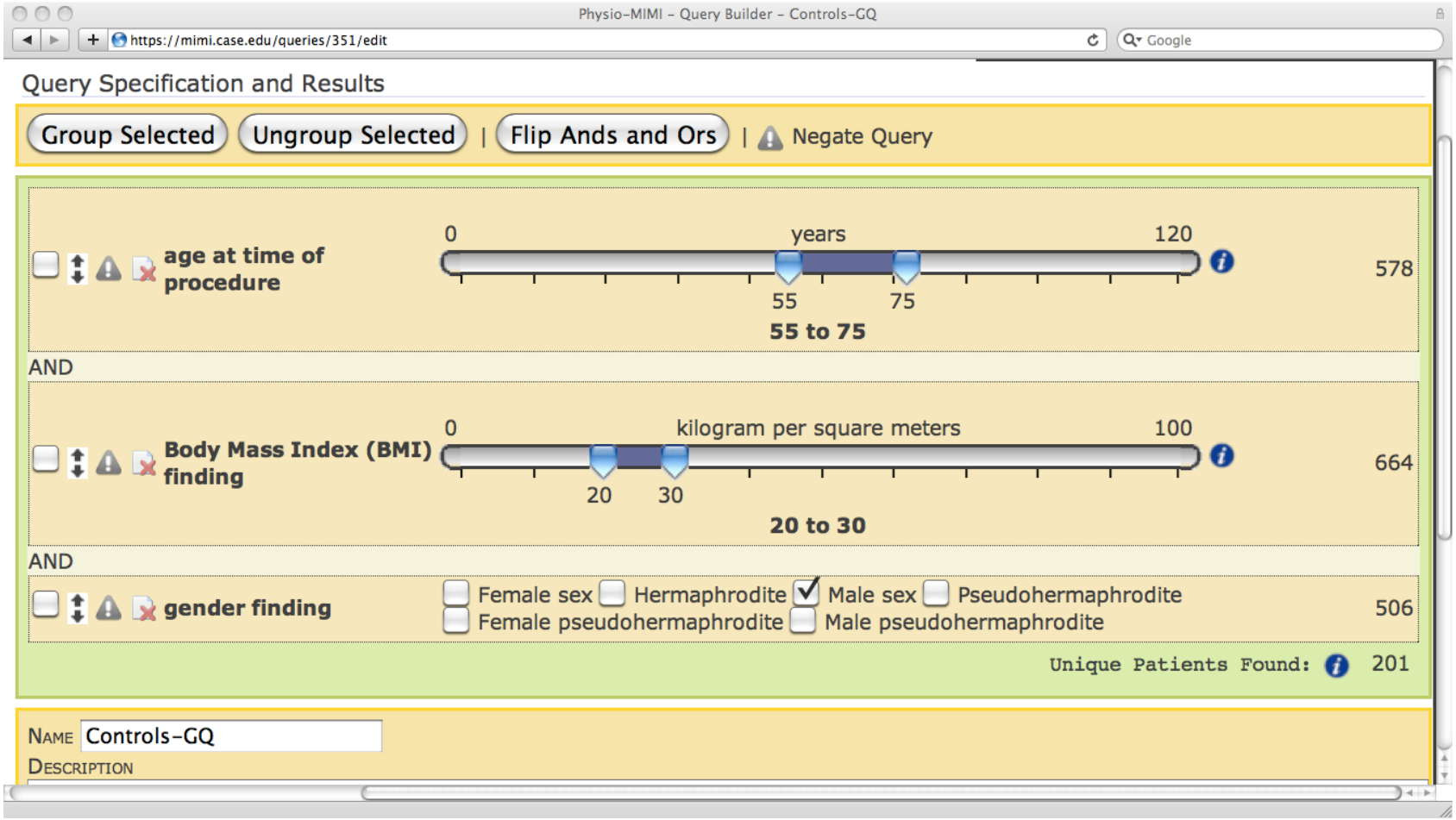
Example (CFS)

- Suppose we are interested in the question of whether sleep parameters (EEG) differ by obesity in age and race matched males
- **Case:** adult 55-75, male, BMI 35-50 (obese)
- **Control:** adult 55-75, male, BMI 20-30 (non-obese)
- **Matching** 1:2 on race (minimize race as a factor initially)

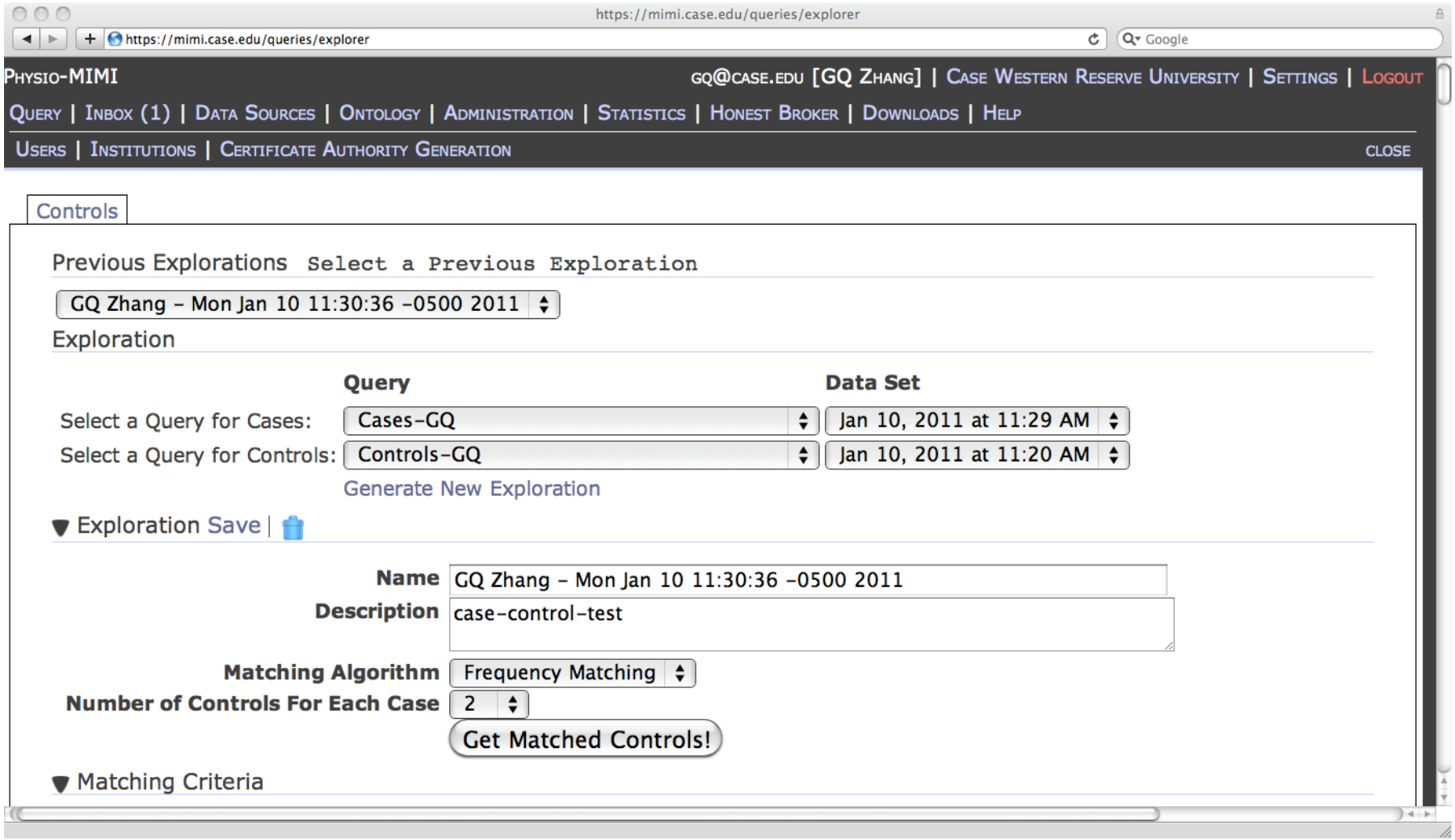
Adult 55-75, male, BMI 35-50



Adult 55-75, male, BMI 20-30

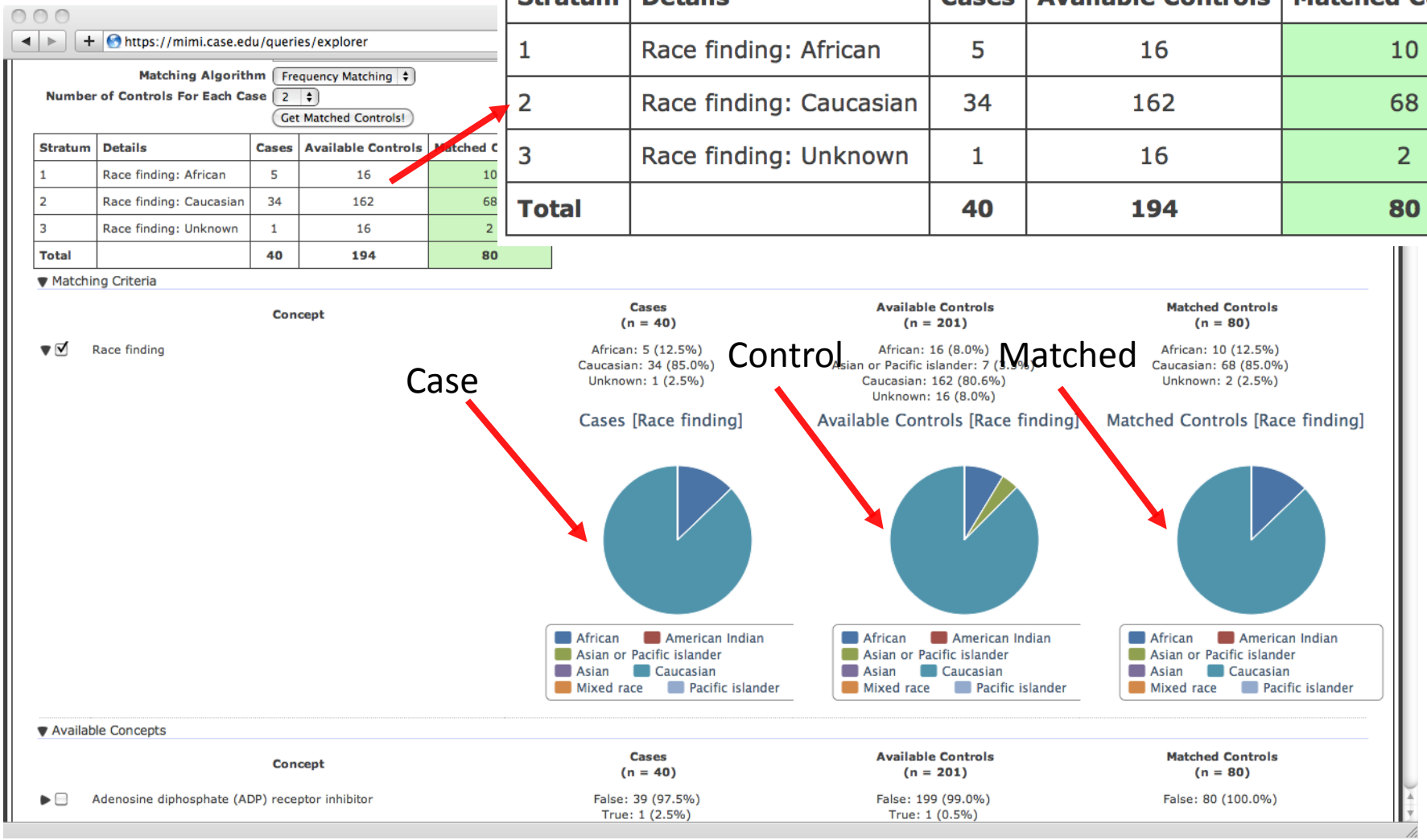


Set up 1:2 Matching

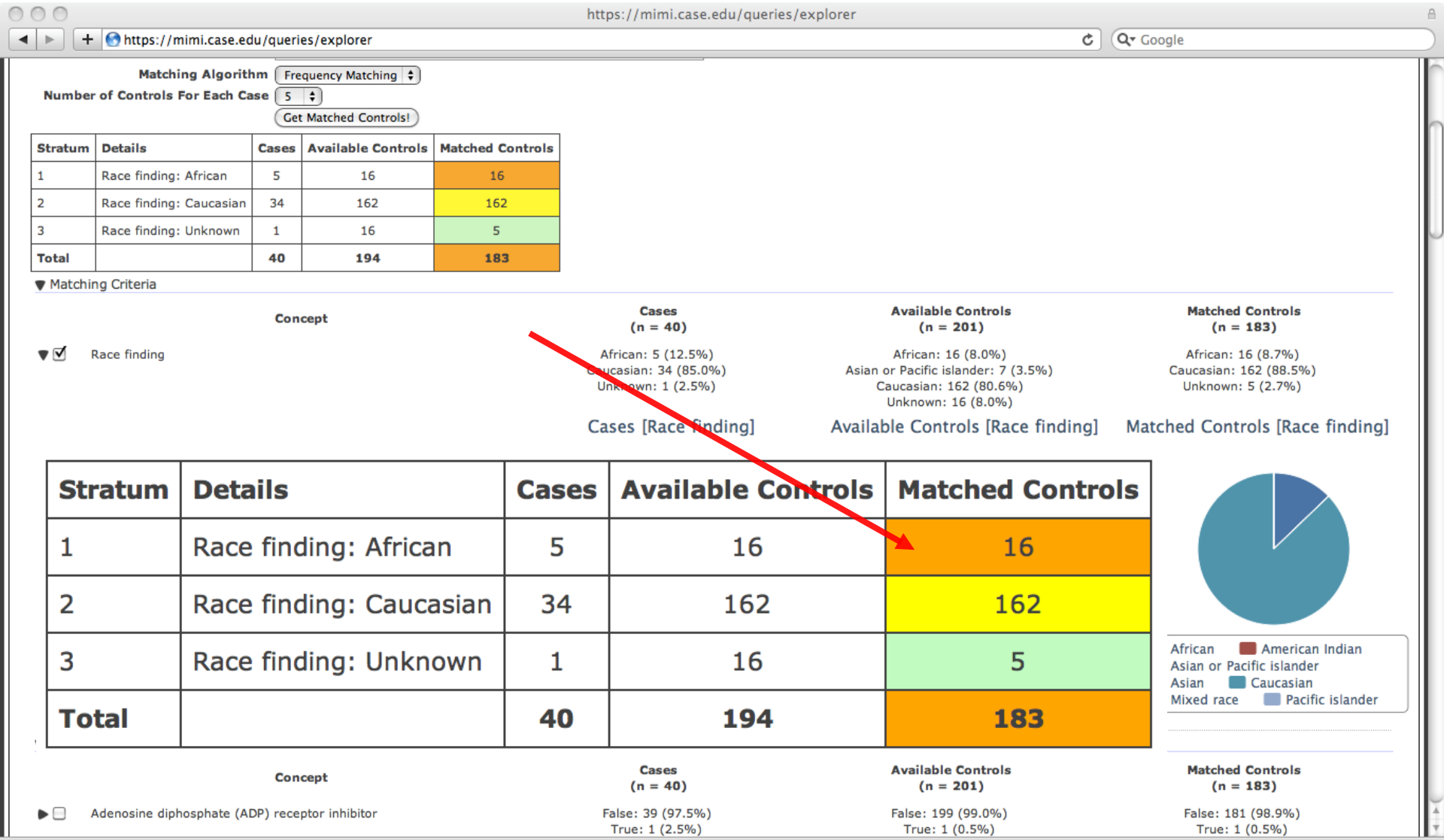


1:2 Matching Result

Stratum	Details	Cases	Available Controls	Matched Controls
1	Race finding: African	5	16	10
2	Race finding: Caucasian	34	162	68
3	Race finding: Unknown	1	16	2
Total		**40**	**194**	**80**

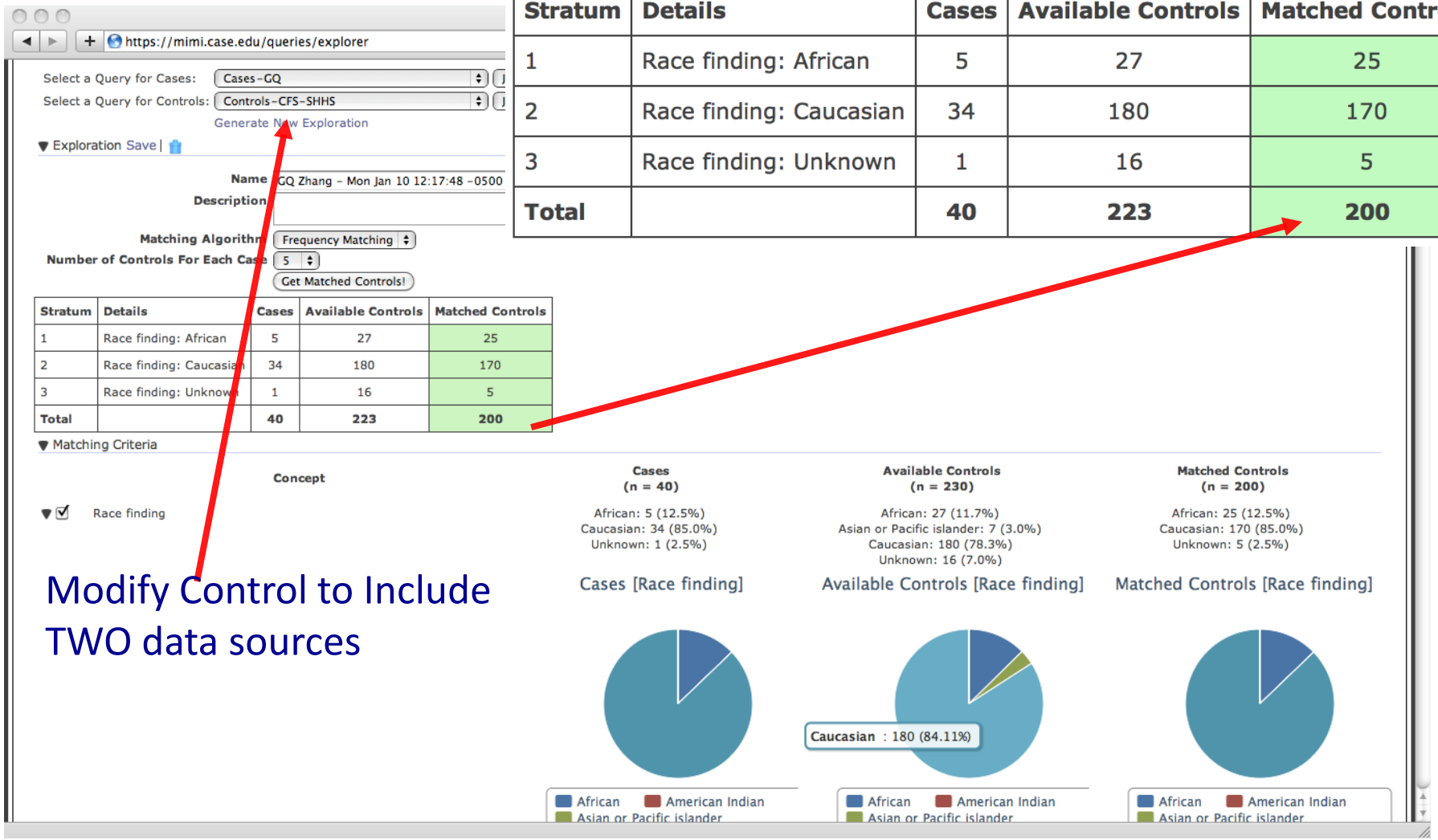


1:5 Matching?


https://mimi.case.edu/queries/explorer
https://mimi.case.edu/queries/explorer
Google
Matching Algorithm
Frequency Matching
Number of Controls For Each Case
5
Get Matched Controls!
Stratum | Details | Cases | Available Controls | Matched Controls
1 | Race finding: African | 5 | 16 | 16
2 | Race finding: Caucasian | 34 | 162 | 162
3 | Race finding: Unknown | 1 | 16 | 5
Total | | 40 | 194 | 183
Matching Criteria
Concept
Cases (n = 40)
Available Controls (n = 201)
Matched Controls (n = 183)
Race finding
African: 5 (12.5%)
Caucasian: 34 (85.0%)
Unknown: 1 (2.5%)
African: 16 (8.0%)
Asian or Pacific islander: 7 (3.5%)
Caucasian: 162 (80.6%)
Unknown: 16 (8.0%)
African: 16 (8.7%)
Caucasian: 162 (88.5%)
Unknown: 5 (2.7%)
Cases [Race finding]
Available Controls [Race finding]
Matched Controls [Race finding]
Stratum | Details | Cases | Available Controls | Matched Controls
1 | Race finding: African | 5 | 16 | 16
2 | Race finding: Caucasian | 34 | 162 | 162
3 | Race finding: Unknown | 1 | 16 | 5
Total | | 40 | 194 | 183
African
American Indian
Asian or Pacific islander
Asian
Caucasian
Mixed race
Pacific islander
Concept
Cases (n = 40)
Available Controls (n = 201)
Matched Controls (n = 183)
Adenosine diphosphate (ADP) receptor inhibitor
False: 39 (97.5%)
True: 1 (2.5%)
False: 199 (99.0%)
True: 1 (0.5%)
False: 181 (98.9%)
True: 1 (0.5%)

1:5 Matching – CFS+SHHS

Stratum	Details	Cases	Available Controls	Matched Controls
1	Race finding: African	5	27	25
2	Race finding: Caucasian	34	180	170
3	Race finding: Unknown	1	16	5
Total		**40**	**223**	**200**



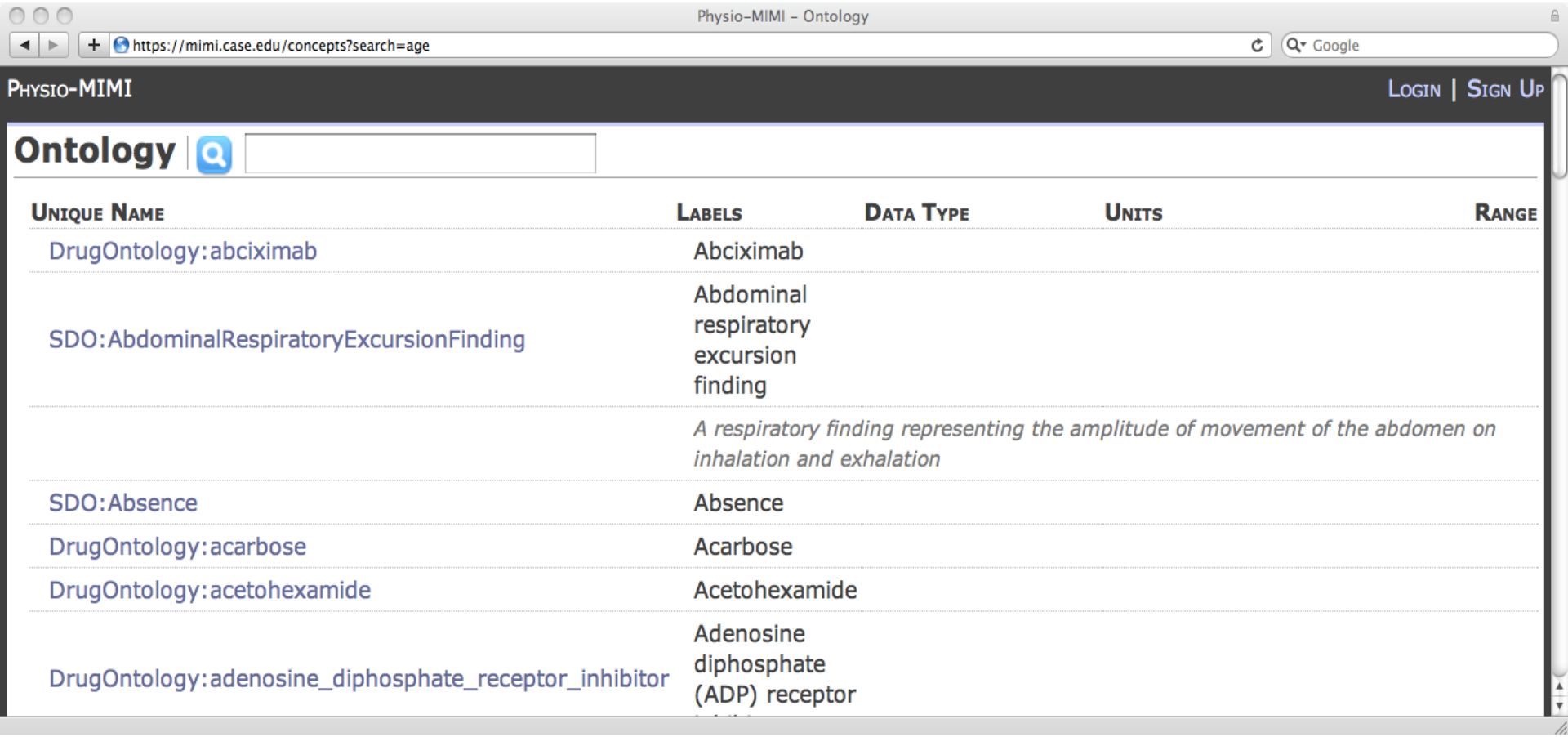
Modify Control to Include TWO data sources

Sleep Domain Ontology (SDO)

- Standardize terminology and semantics (define variations) [RO]
- Facilitate definition of data elements
- Valuable for data collection, data curation
- Data integration
- Data sharing and access
- Take advantage of progress in related areas (e.g. Gene Ontology)
- Improving data quality – provenance, reproducibility

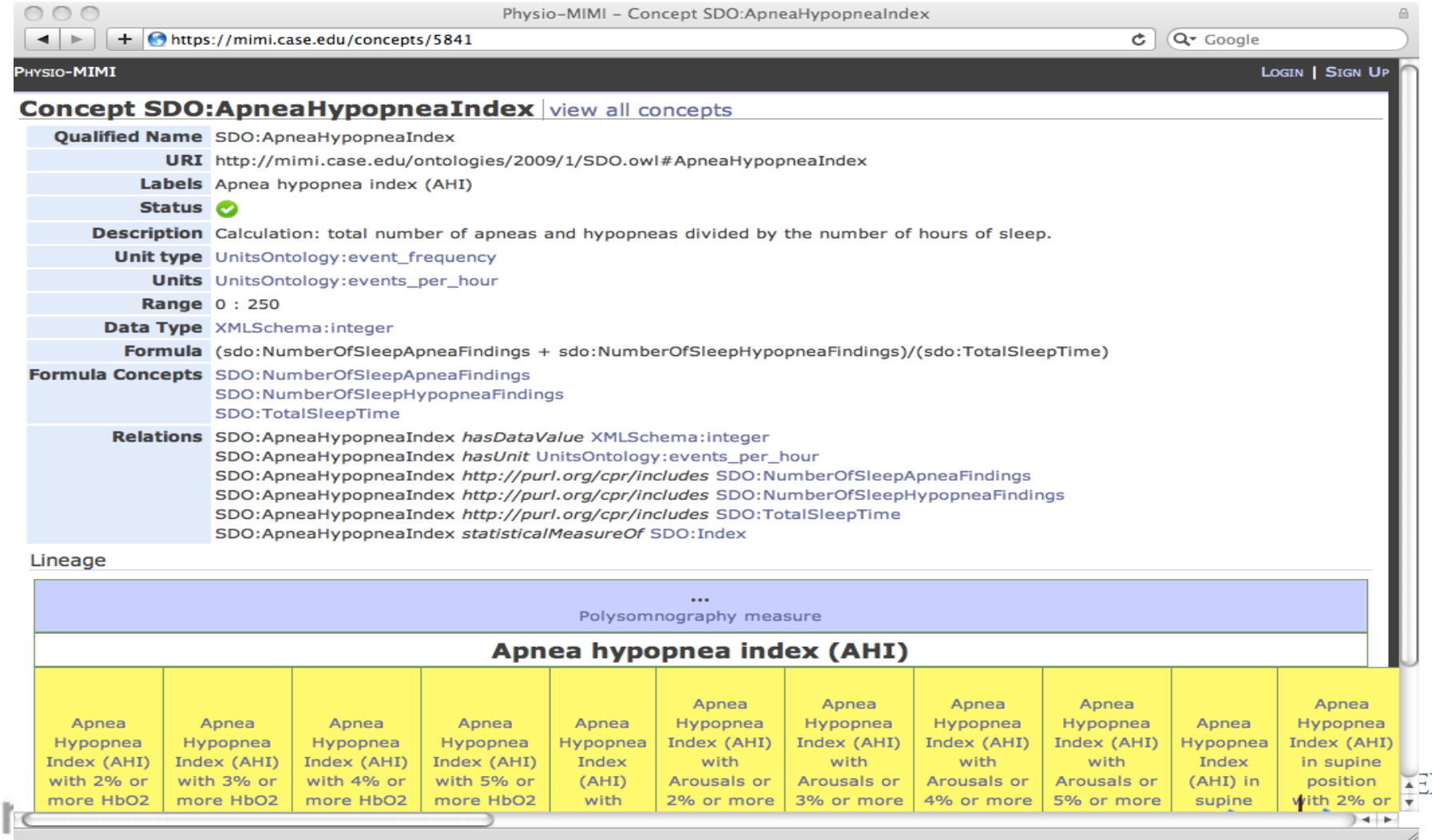
Sleep Domain Ontology (SDO)

https://mimi.case.edu/concepts

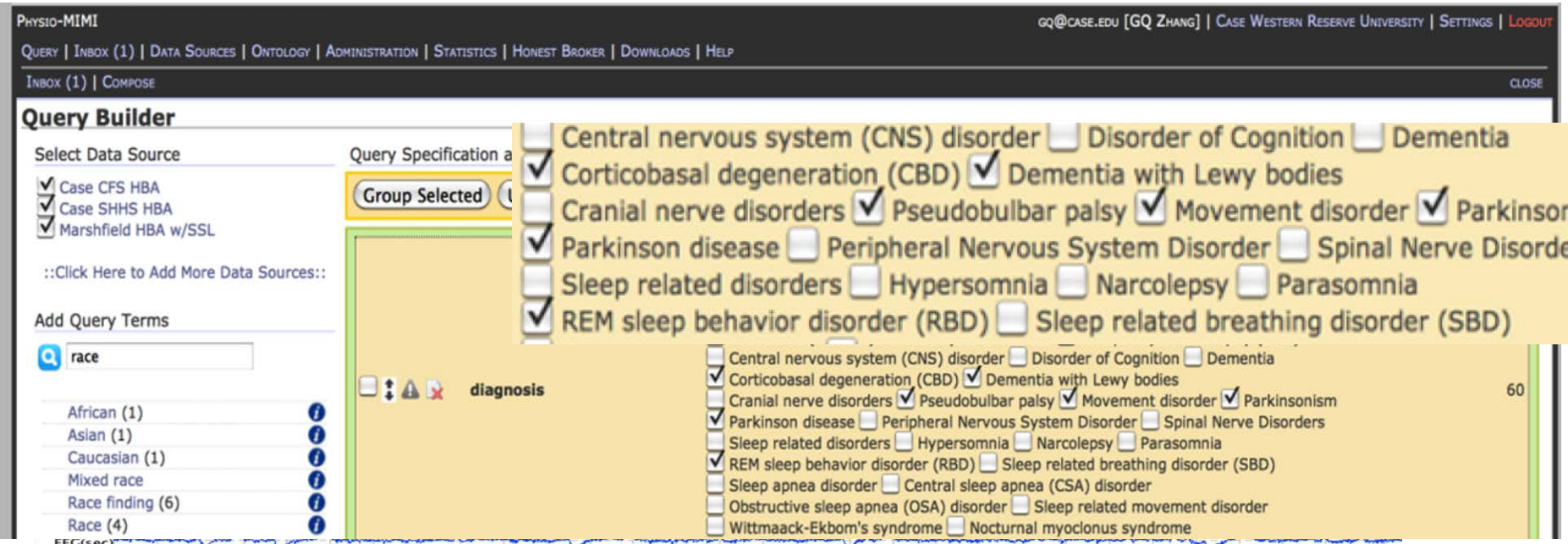


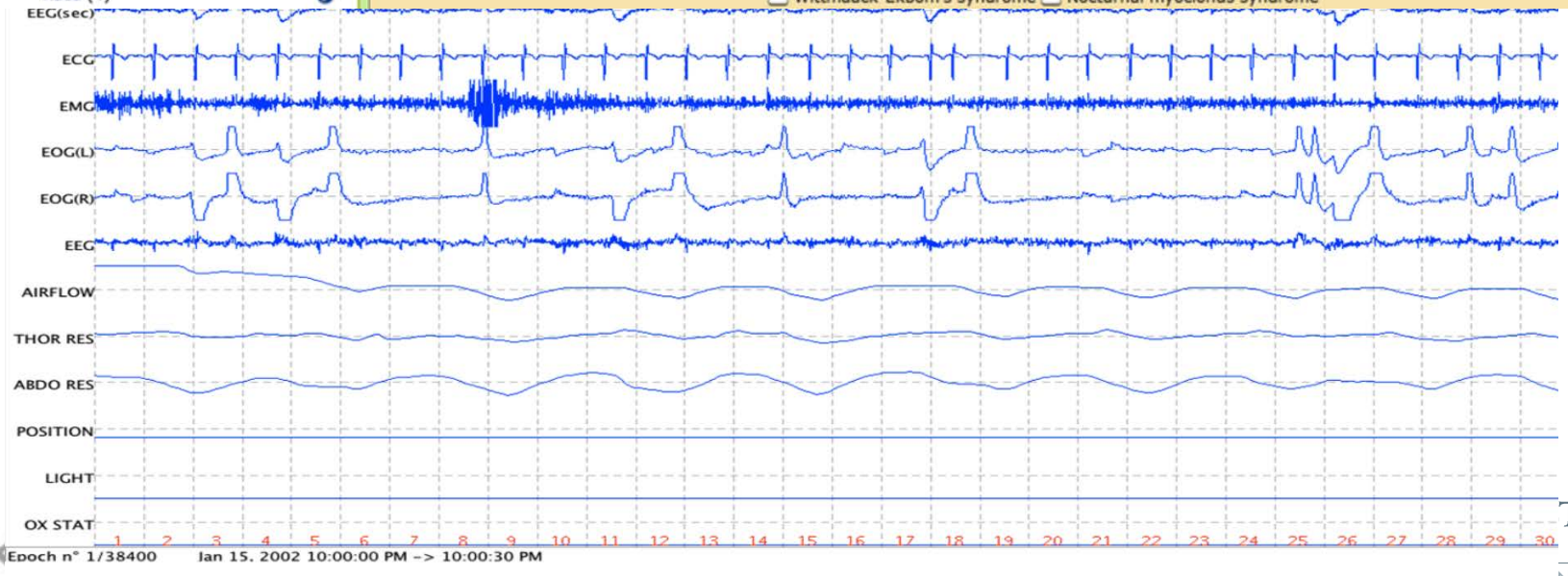
Sleep Domain Ontology (SDO)

https://mimi.case.edu/concepts



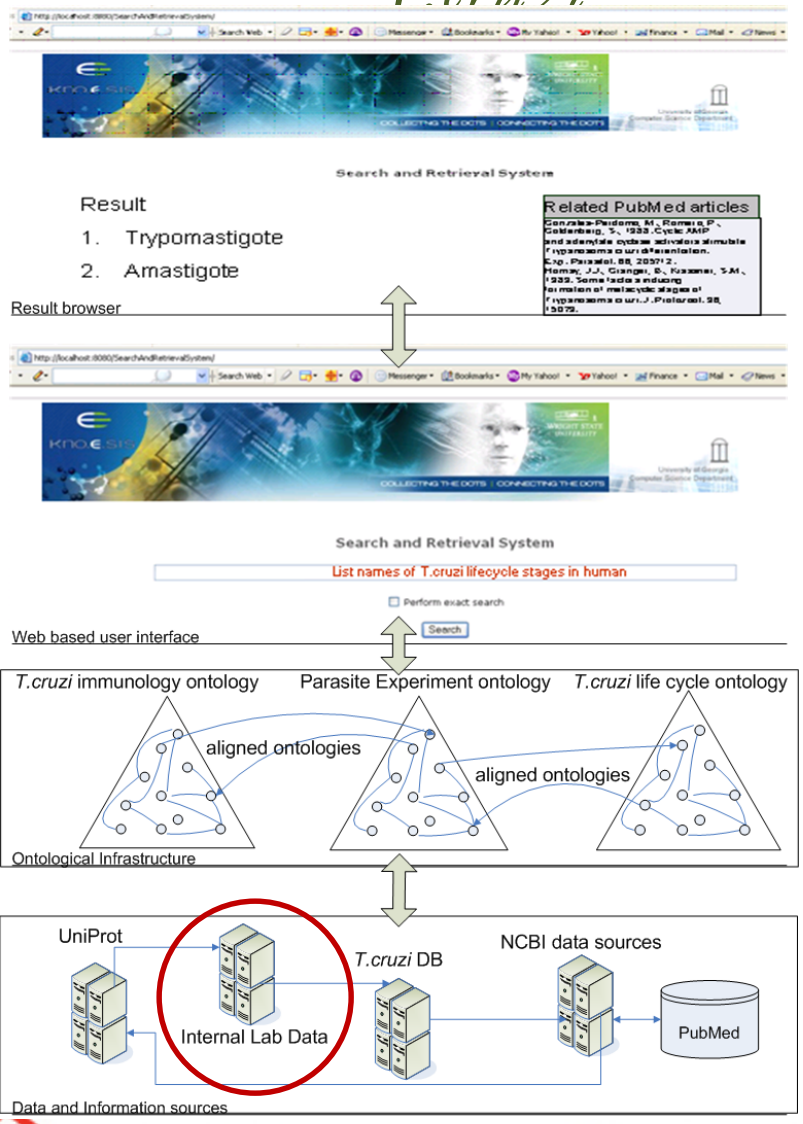
VISAGE Query Builder showing a data query on Parkinsonian Disorders and REM sleep behavior disorder with race demographics





Semantic Web + Provenance + Bench Research =
T.cruzi Semantic Problem Solving Environment

Semantic Problem Solving Environment for *T.cruzi*



THE UNIVERSITY OF GEORGIA

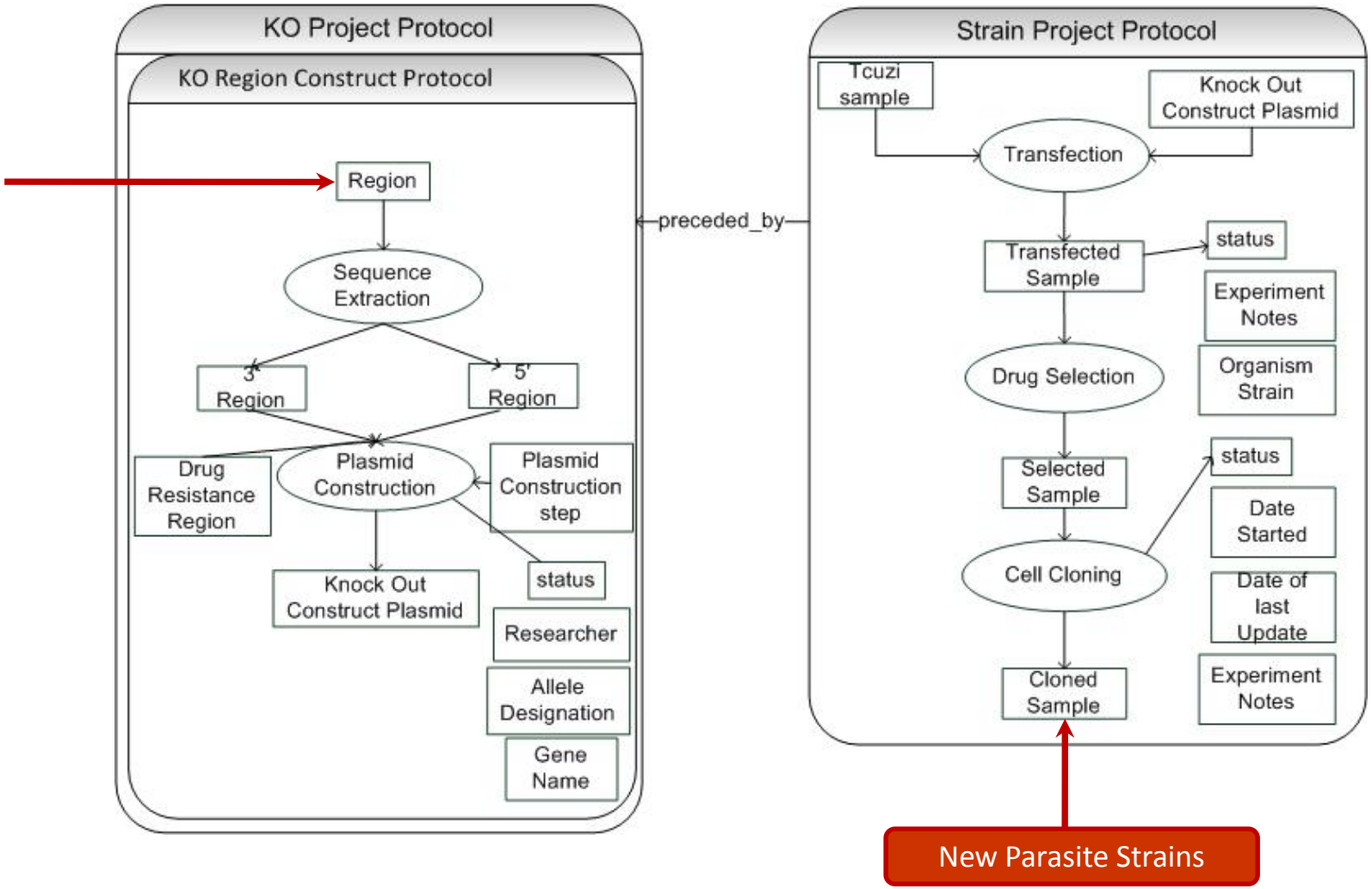
STANFORD UNIVERSITY

National Heart Lung and Blood Institute

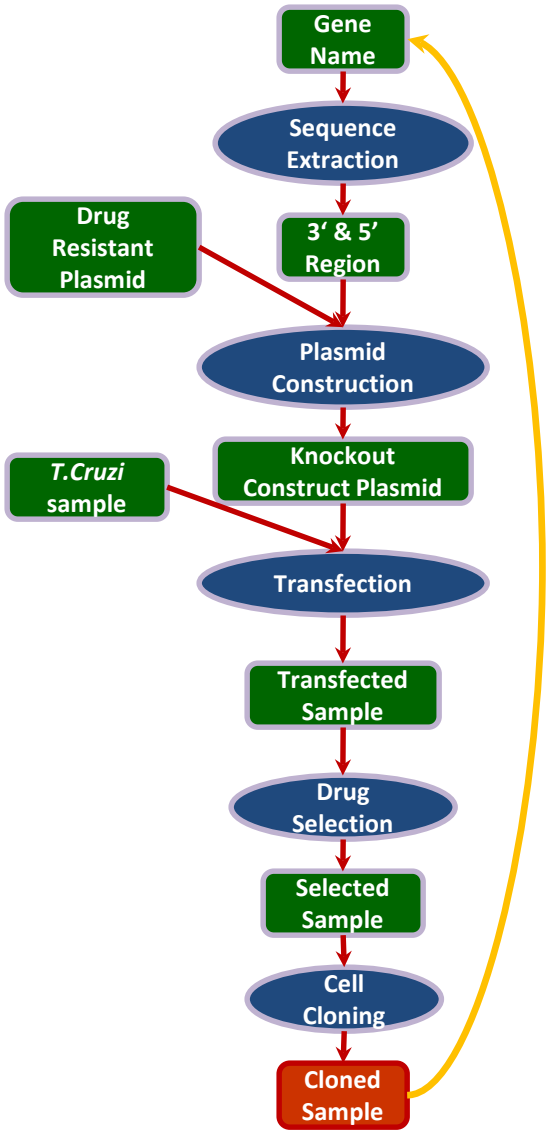
kno.e.sis

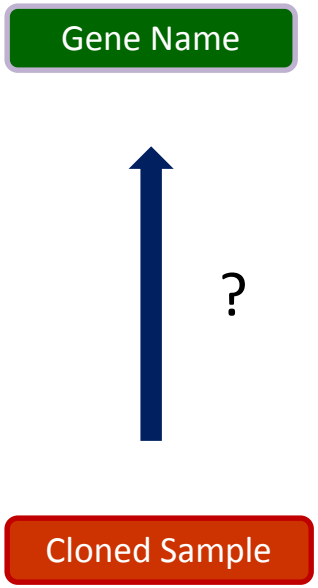
CASE WESTERN RESERVE UNIVERSITY
SCHOOL OF MEDICINE

Provenance in Scientific Experiments

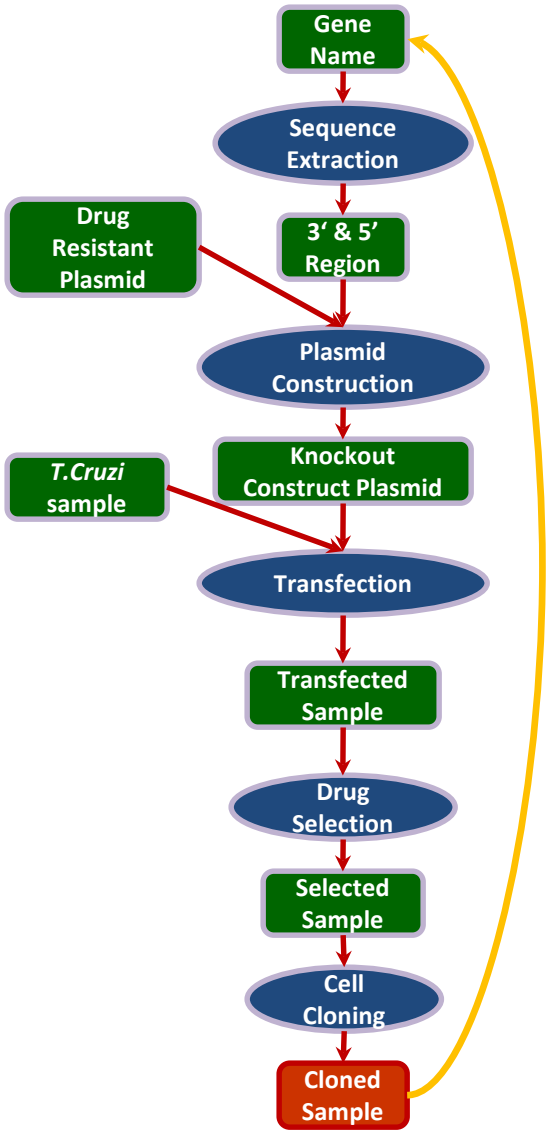


Provenance in Scientific Experiments



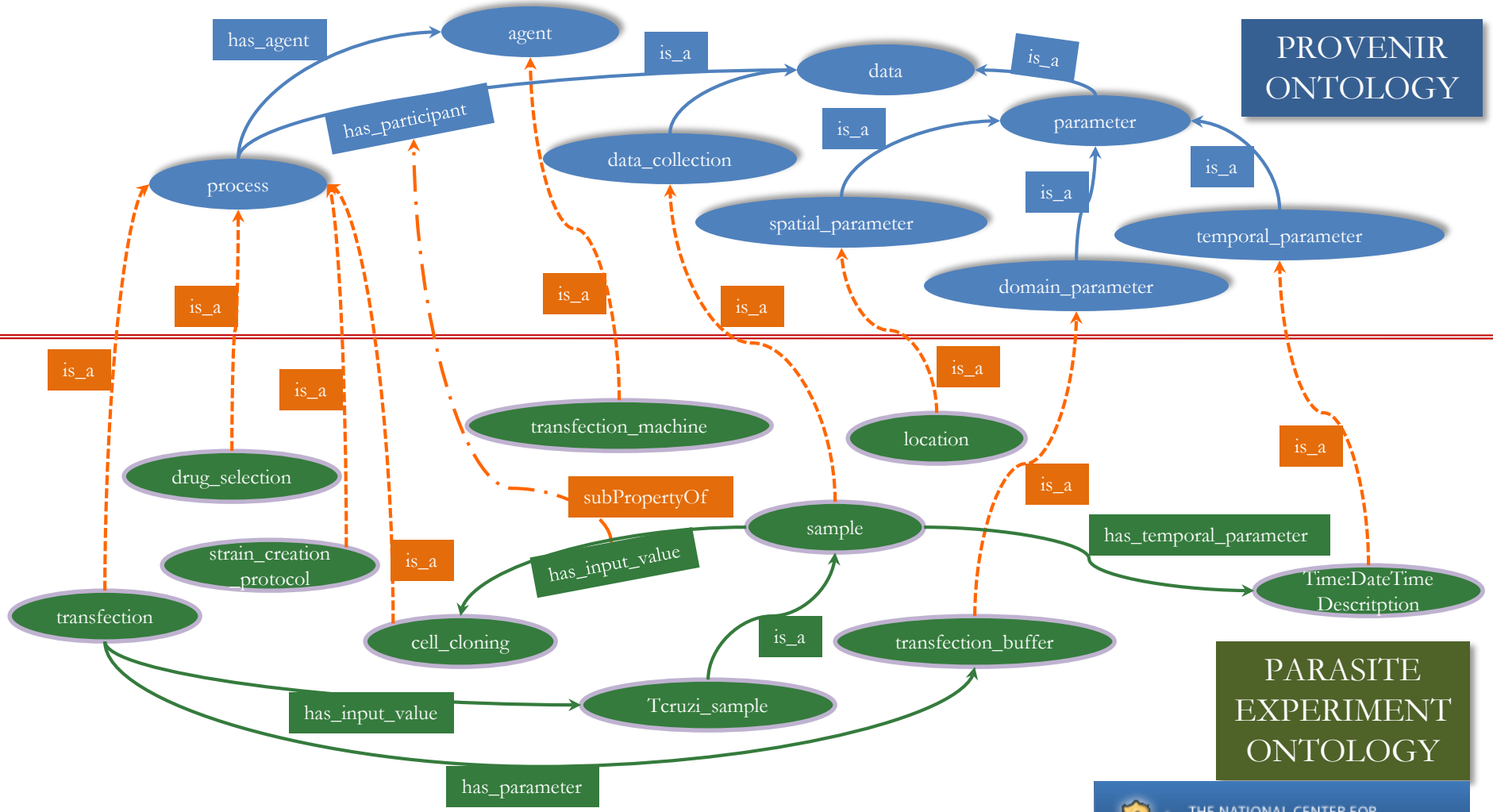


Provenance in Scientific Experiments



- Provenance from the French word "*provenir*" describes the lineage or history of a data entity
- For Verification and Validation of Data Integrity, Process Quality, and Trust
- Semantic Provenance Framework addresses three aspects [Prov]
  - Provenance Modeling
  - Provenance Query Infrastructure
  - Scalable Provenance System

Domain-specific Provenance ontology



- Total Number of Classes - 118
- DL Expressivity – ALCHQ(D)

Provenance Query Classification

Classified Provenance Queries into Three Categories

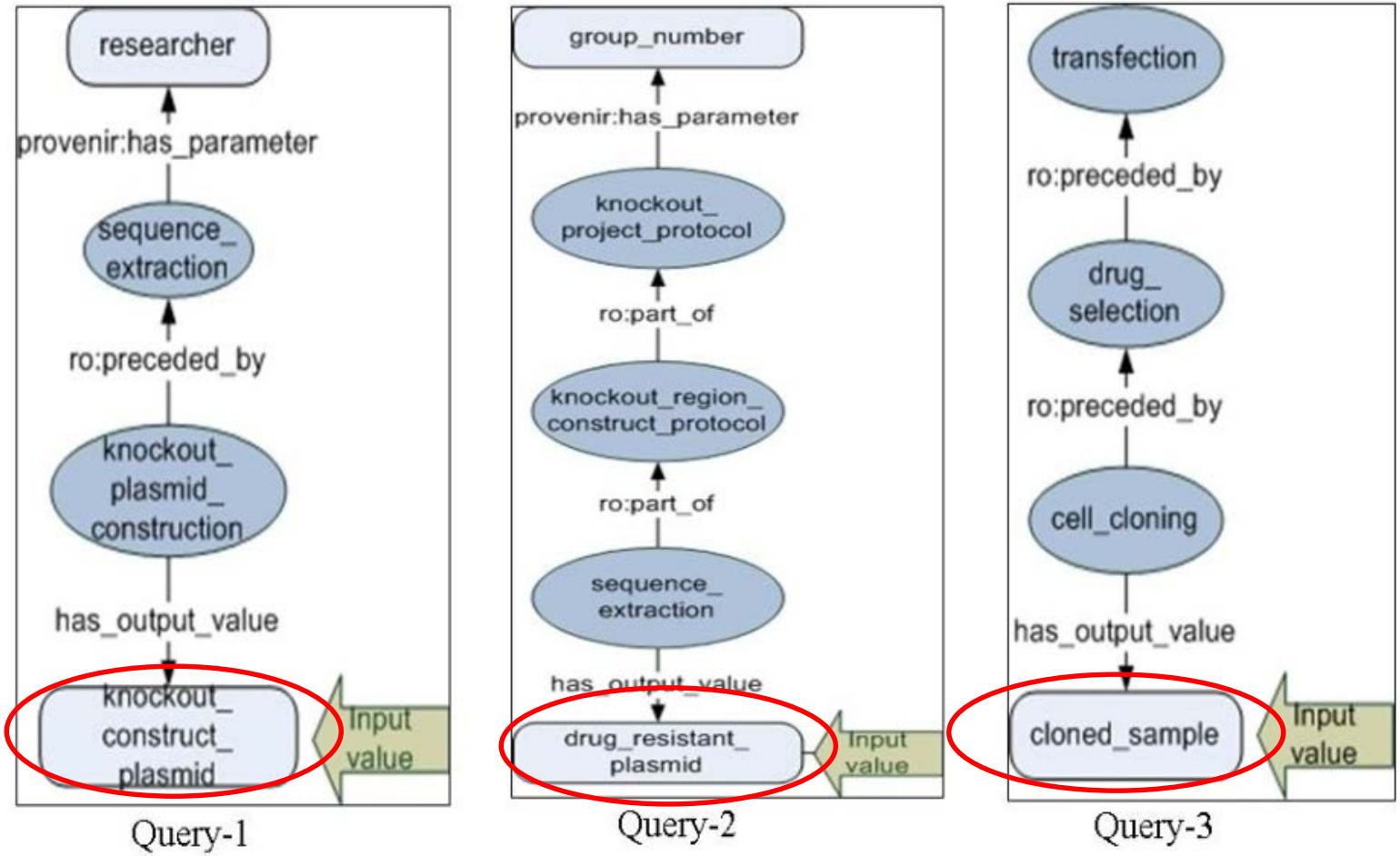
- Type 1: Querying for Provenance Metadata
  - Example: *Which gene was used create the cloned sample with ID = 66?*
- Type 2: Querying for Specific Data Set
  - Example: *Find all knockout construct plasmids created by researcher Michelle using "Hygromycin" drug resistant plasmid between April 25, 2008 and August 15, 2008*
- Type 3: Operations on Provenance Metadata
  - Example: Were the two cloned samples 65 and 46 prepared under similar conditions – compare the associated provenance information

Provenance Query Operators

Four Query Operators – based on Query Classification

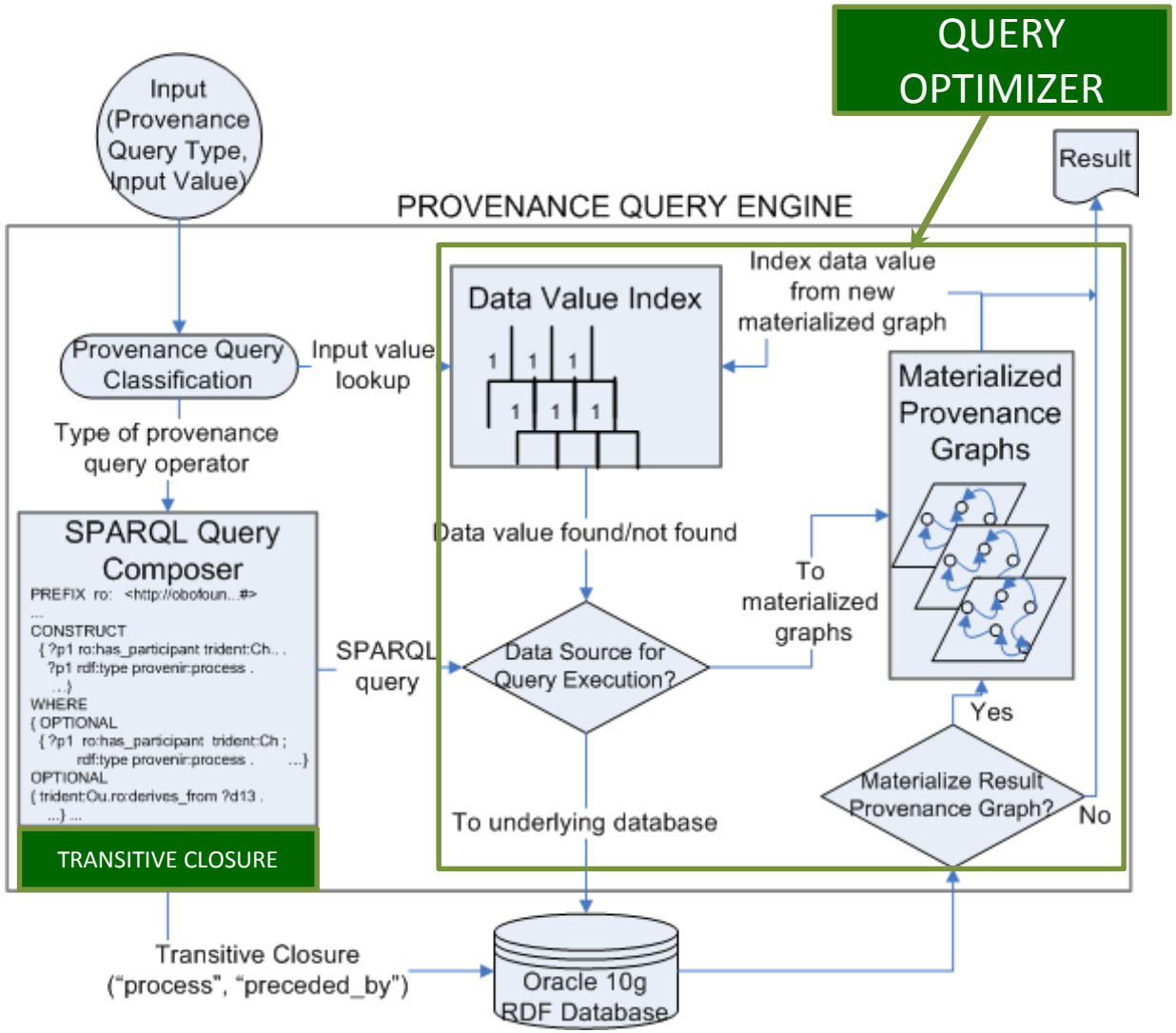
- *provenance* () – Closure operation, returns the complete set of provenance metadata for input data entity
- *provenance_context*() - Given set of constraints defined on provenance, retrieves datasets that satisfy constraints
- *provenance_compare* () - adapt the RDF graph equivalence definition
- *provenance_merge* () - Two sets of provenance information are combined using the RDF graph merge

Answering Provenance Queries using *provenance* () Operator

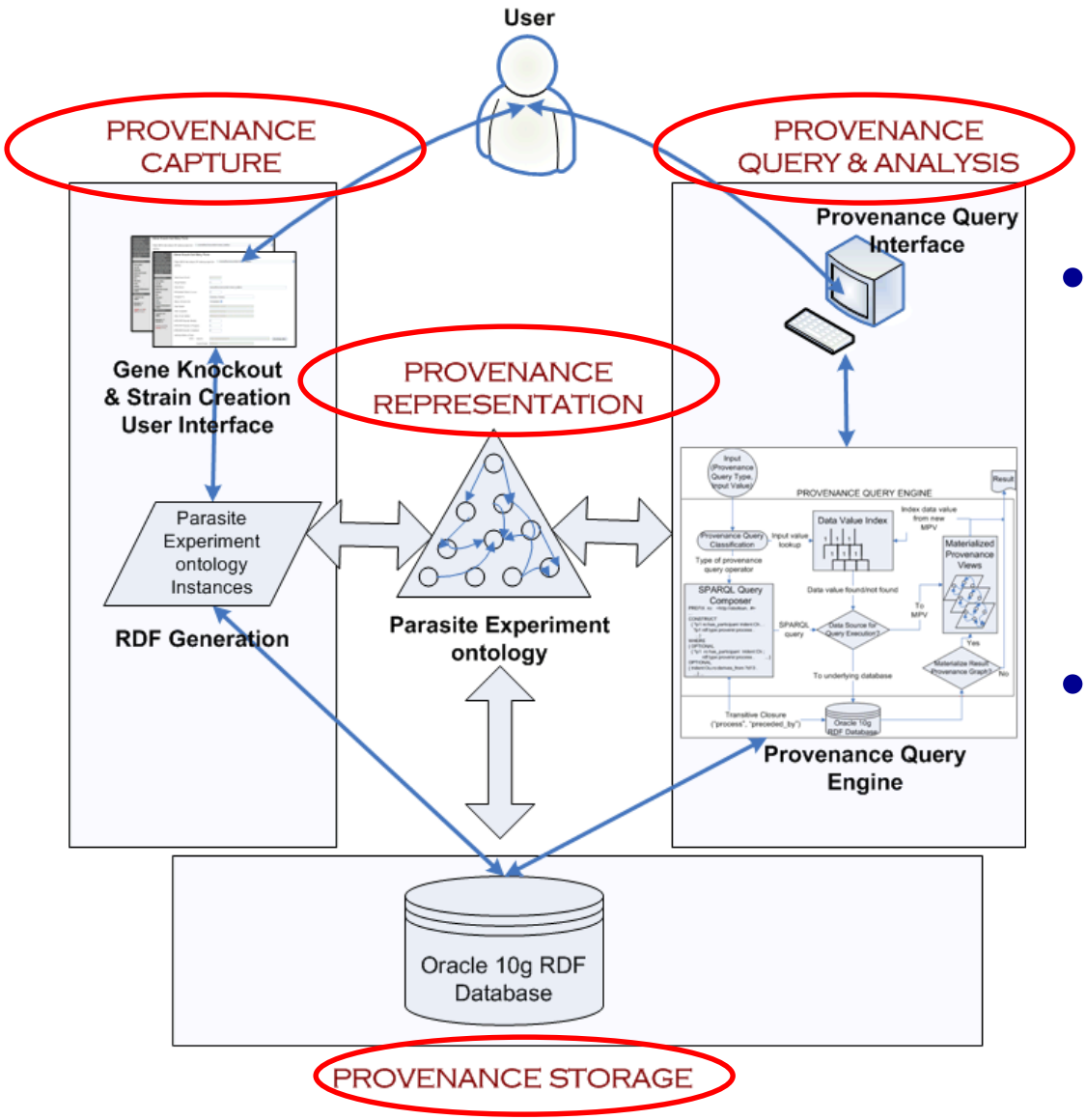


Implementation: Provenance Query Engine

- Three modules:
  - Query Composer
  - Transitive closure
  - Query Optimizer
- Deployable over a RDF store with support for reasoning



Application in *T.cruzi* SPSE Project



- Provenance tracking for gene knockout, strain creation, proteomics, microarray experiments
- Part of the Parasite Knowledge Repository [BKR]

W3C Provenance Working Group

- Define a "provenance interchange language for publishing and accessing provenance"
- Three working drafts:
  - PROV-Data Model: A conceptual model for provenance representation
  - PROV-Ontology: An OWL ontology for provenance representation
  - PROV-Access and Query: A framework to query and retrieve provenance on the Web

Semantic Web + Clinical Practice Informatics =
Active Semantic Electronic Medical Record
(ASEMR)

Semantic Web application in use

In daily use at Athens Heart Center

- 28 person staff
  - Interventional Cardiologists
  - Electrophysiology Cardiologists
- Deployed since January 2006
- 40-60 patients seen daily
- 3000+ active patients
- Serves a population of 250,000 people

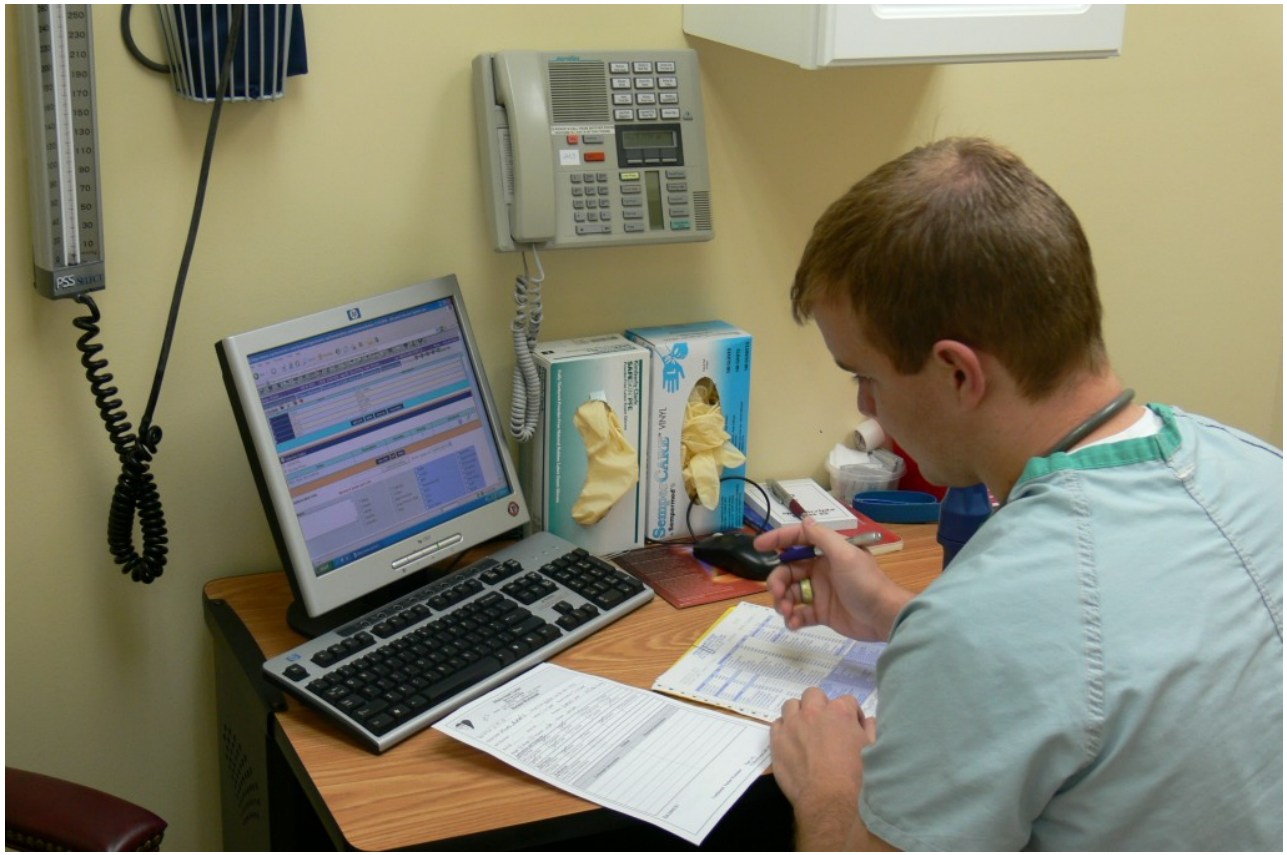
Information Overload in Clinical Practice

- New drugs added to market
  - Adds interactions with current drugs
  - Changes possible procedures to treat an illness
- Insurance Coverage's Change
  - Insurance may pay for drug X but not drug Y even though drug X and Y are equivalent
  - Patient may need a certain diagnosis before some expensive test are run
- Physicians need a system to keep track of ever changing landscape

System though out the practice

System though out the practice

System though out the practice



System though out the practice

Active Semantic Document (ASD)

A document (typically in XML) with the following features:

- Semantic annotations
  - Linking entities found in a document to ontology
  - Linking terms to a specialized lexicon [TR]
- Actionable information
  - Rules over semantic annotations
  - Violated rules can modify the appearance of the document (Show an alert)

Active Semantic Patient Record

- An application of ASD
- Three Ontologies
  - Practice

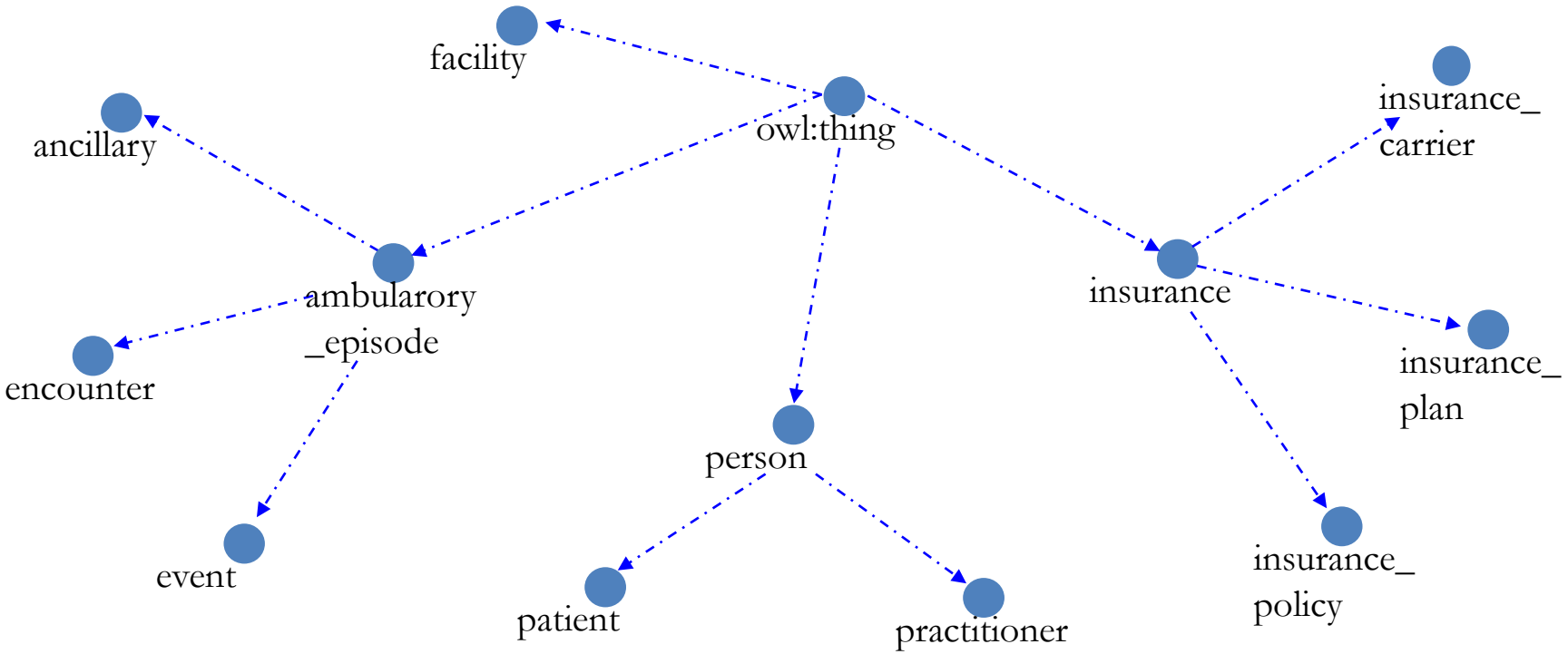
    Information about practice such as patient/physician data
  - Drug

    Information about drugs, interaction, formularies, etc.
  - ICD/CPT

    Describes the relationships between CPT and ICD codes
- Medical Records in XML created from database

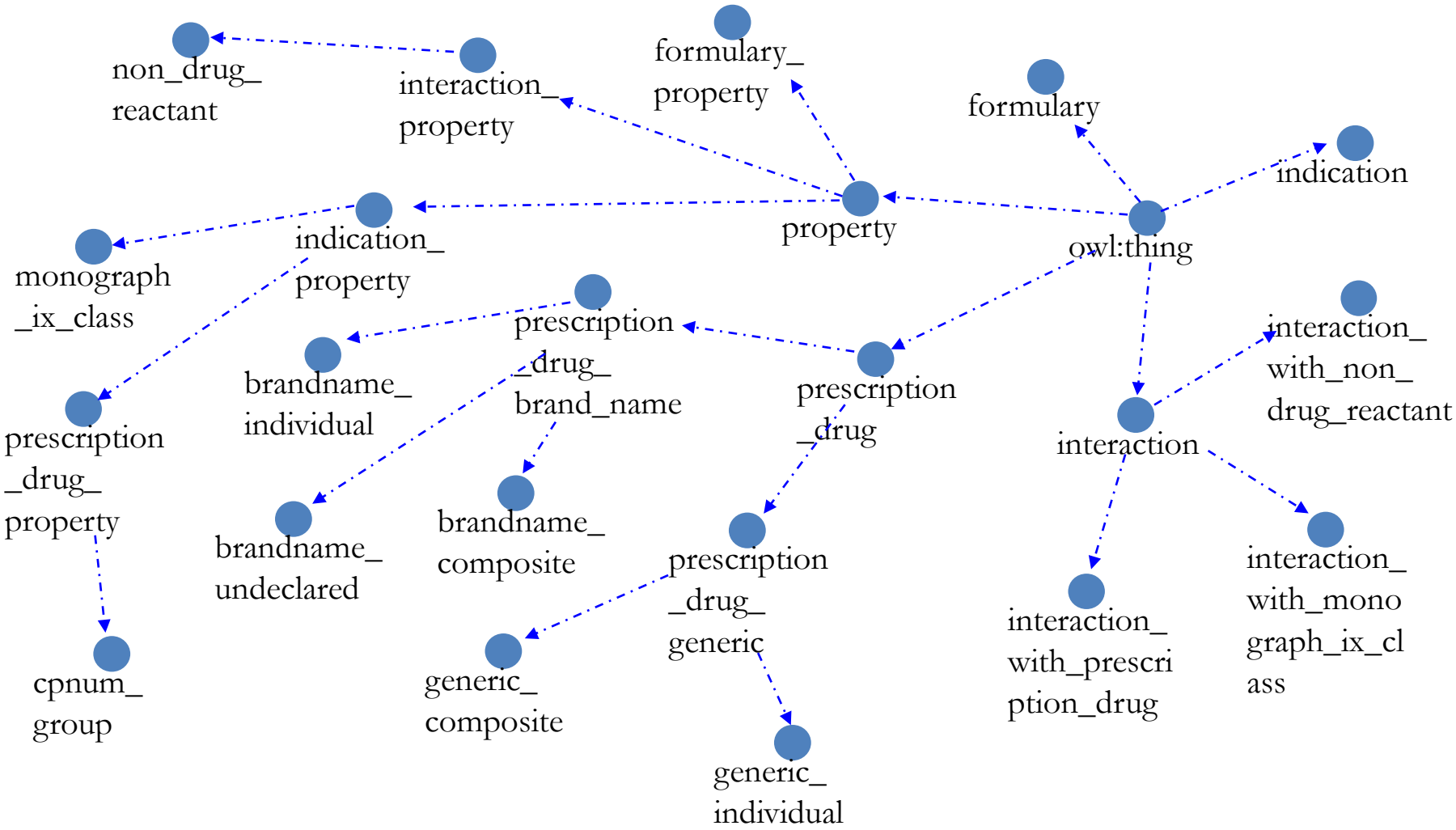
Practice Ontology Hierarchy

(showing is-a relationships)

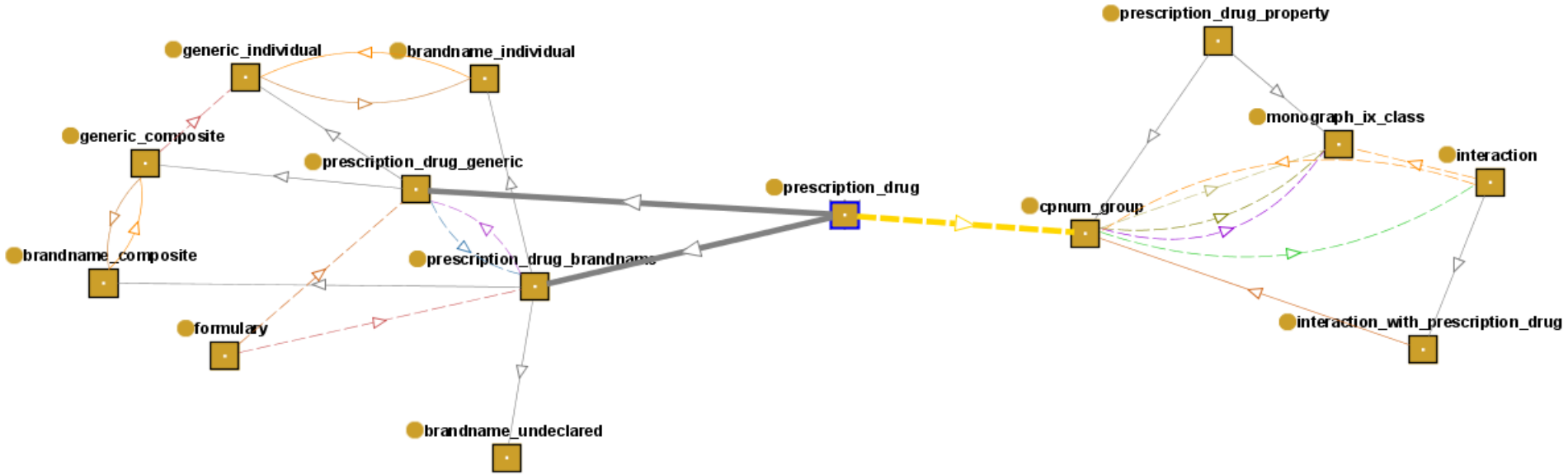


Drug Ontology Hierarchy

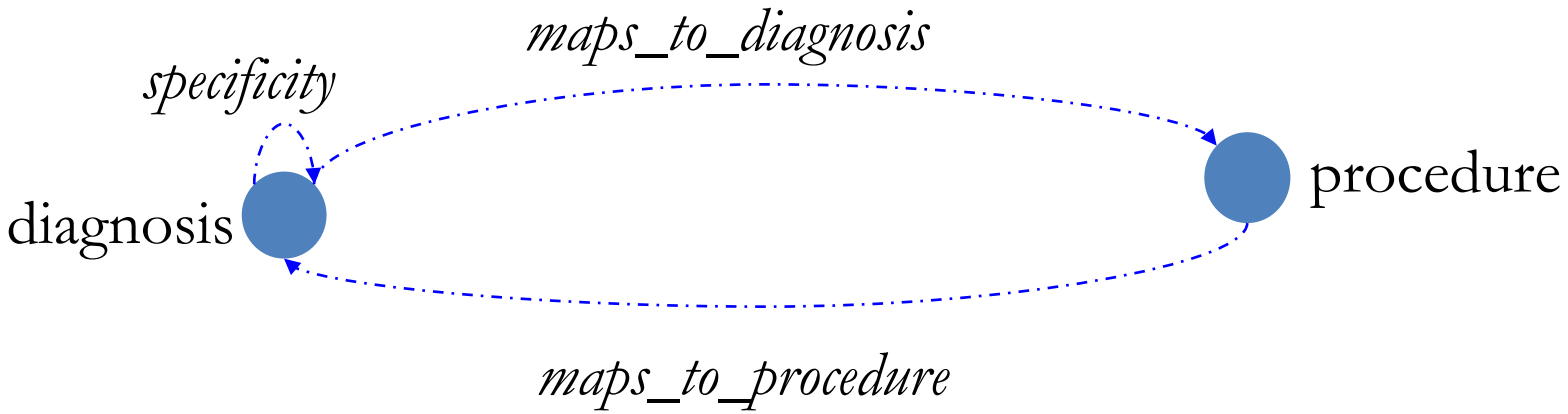
(showing is-a relationships)



Drug Ontology showing neighborhood of *PrescriptionDrug* concept



Part of Procedure/Diagnosis/ICD9/CPT Ontology



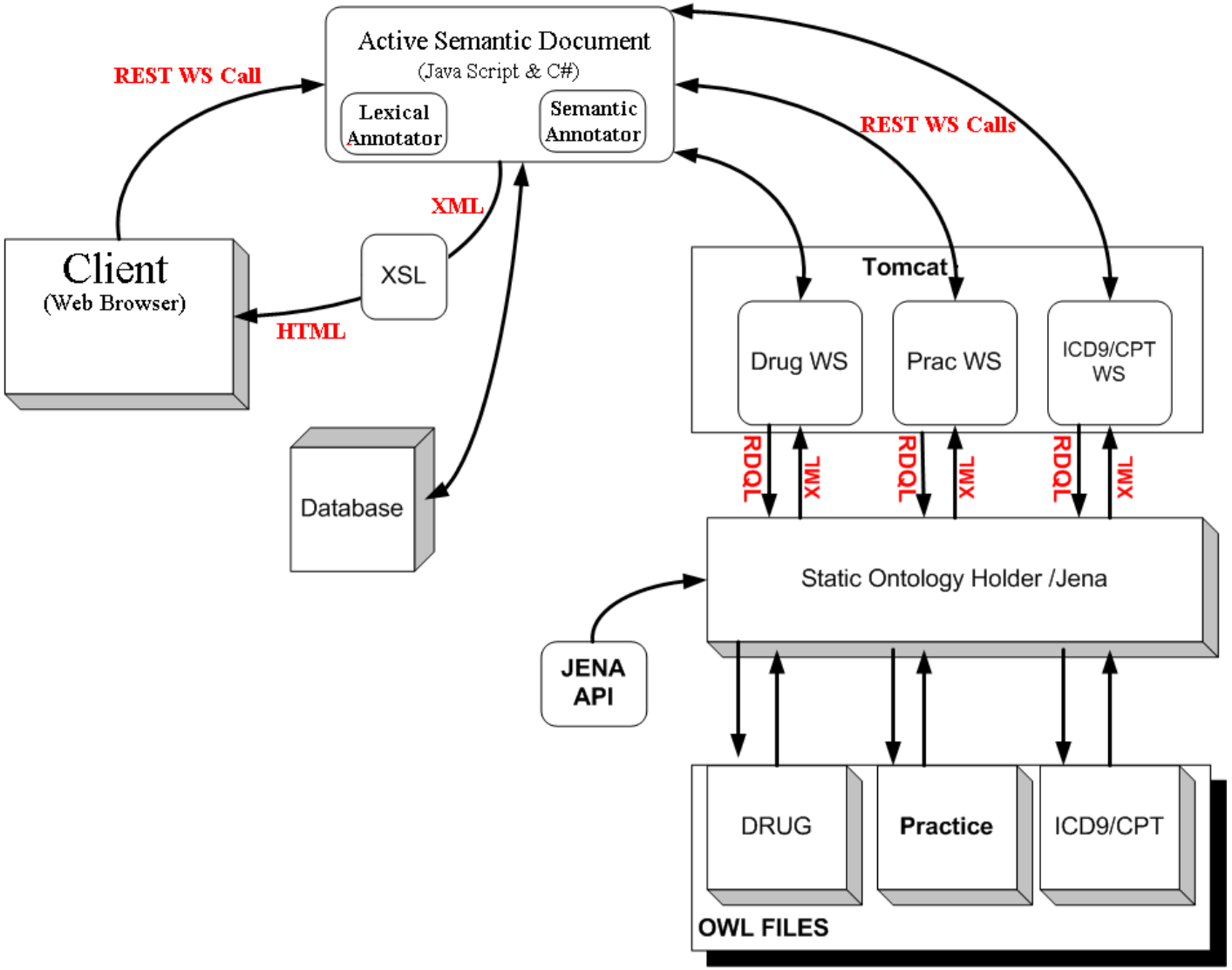
Semantic Technologies in Use

- Semantic Web: OWL, RDF/RDQL, Jena
  - OWL (constraints useful for data consistency), RDF
  - Rules are expressed as RDQL
  - REST Based Web Services: from server side
- Web 2.0: client makes AJAX calls to ontology, also auto complete

Problem:

- Jena main memory- large memory footprint, future scalability challenge
- Using Jena's persistent model (MySQL) noticeably slower

Architecture & Technology



Benefits: Athens Heart Center Practice Growth

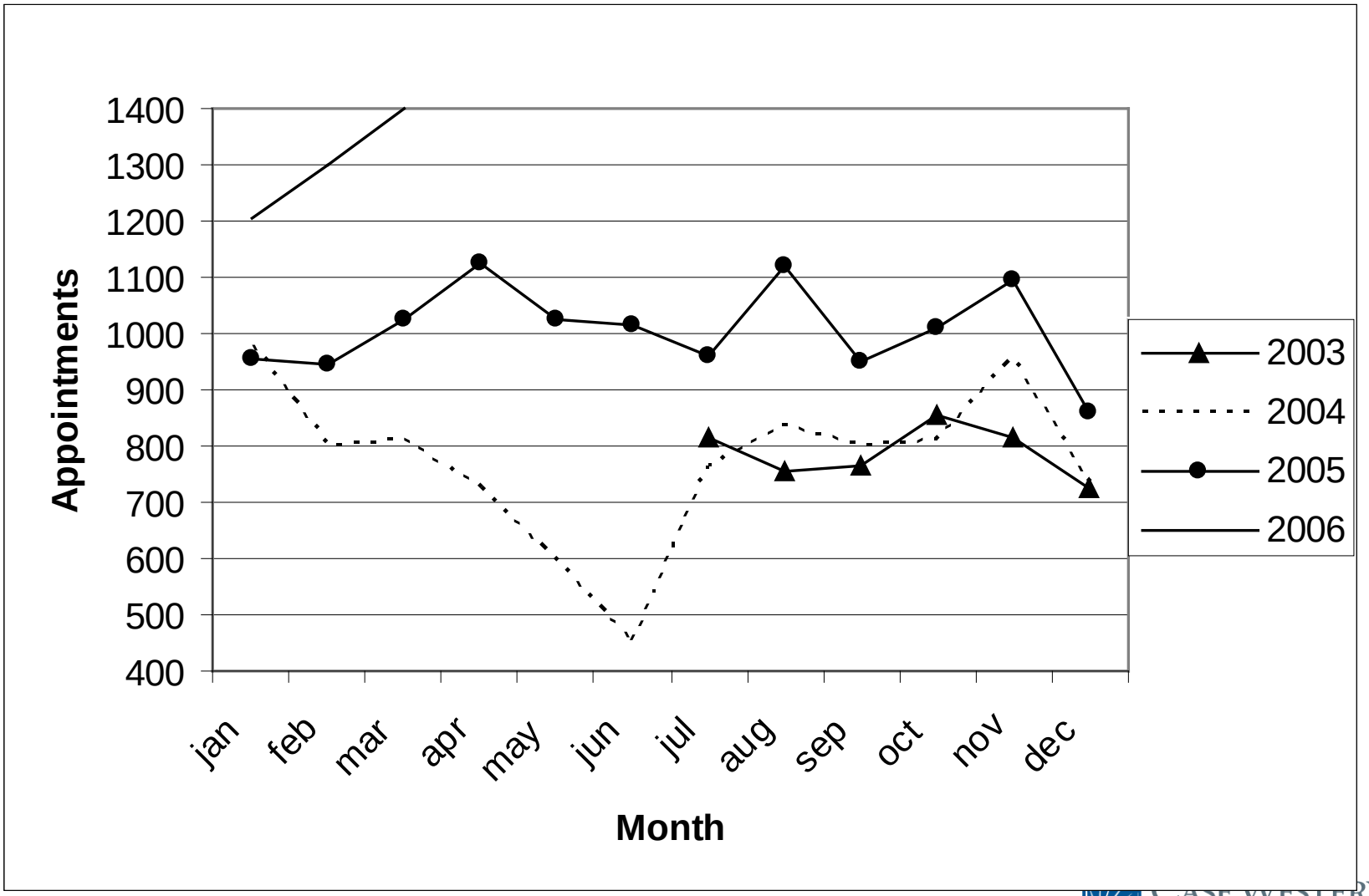


Chart Completion before the preliminary deployment of the ASMER

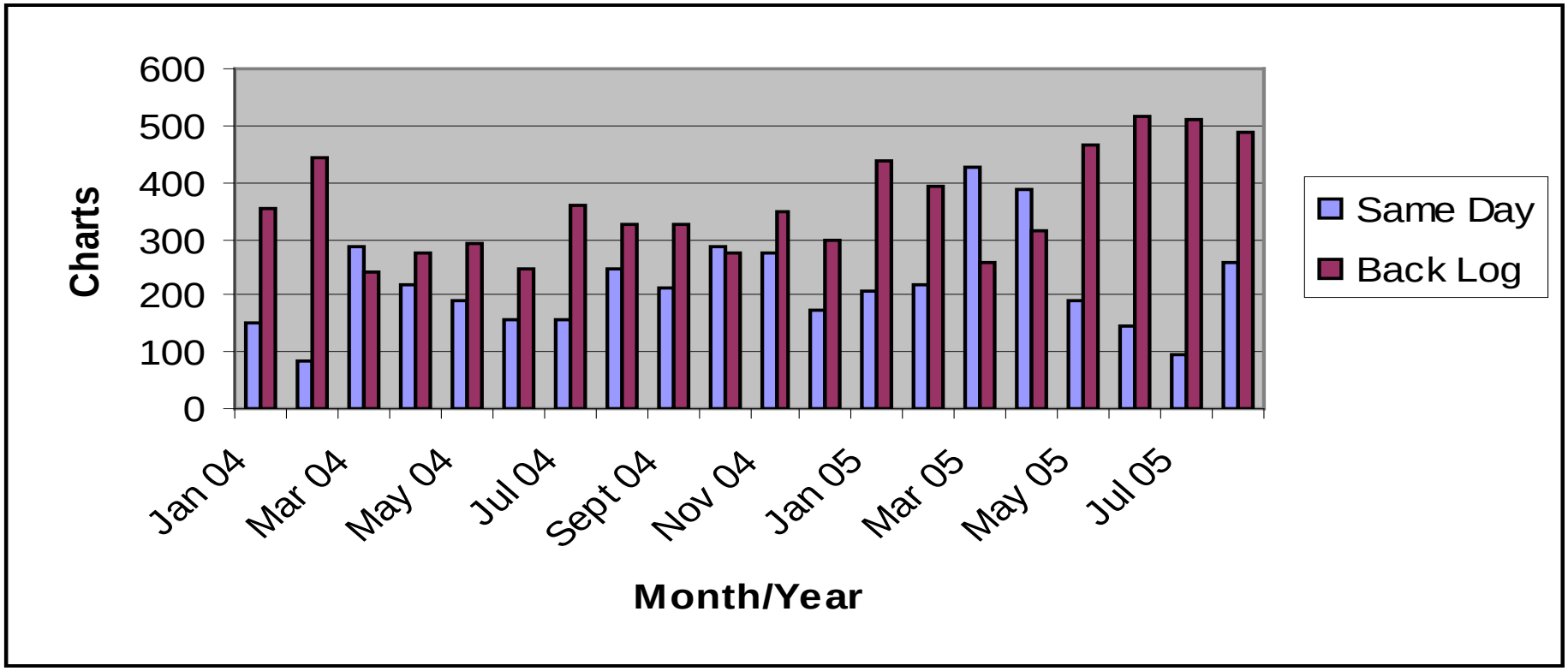
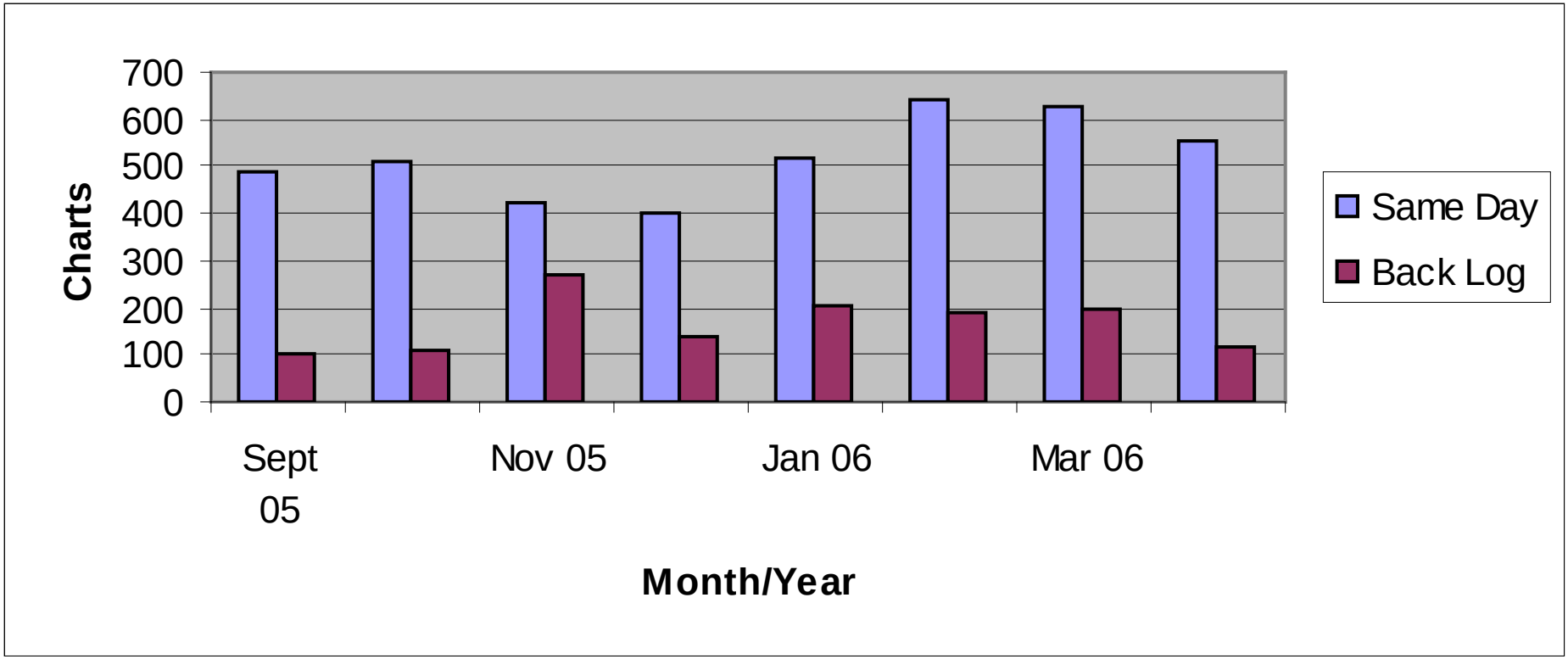


Chart Completion after the preliminary deployment of the ASMER



Benefits of current system

- Error prevention (drug interactions, allergy)
  - Patient care
  - insurance
- Decision Support (formulary, billing)
  - Patient satisfaction
  - Reimbursement
- Efficiency/time
  - Real-time chart completion
  - "semantic" and automated linking with billing

Demo

On-line demo of Active Semantic Electronic Medical Record

deployed and in use at Athens Heart Center

Challenges, Opportunities, and Future Direction

Conclusions

Benefits of SW in Health Informatics:

- RDF a "universal" data model; Application-purpose agnostic (clinical care vs research)
- Integration "ready," supporting distributed query out of box
- Semantic interoperability addressed at root level
- Better support of user interfaces for data capture, data query, data integration
- Scalability demonstrated

Challenges and Future Directions

- Design and implementation of health information systems with RDF as primary data store from ground up
- User-friendly graphical query interface on top of SPARQL
- Managing Protected Health Information (PHI) e.g. data encryption "at rest" for RDF store
- From retrospective annotation of data (with ontology) to prospective annotation of data: ontology-driven data capture with annotation happening at the point of primary source (eliminating the need to annotate data retrospectively)
- Let ontology drive "everything"

References

- [RDF] Manola F, Miller, E.(Eds.). RDF Primer. 2004; Available from: http://www.w3.org/TR/rdf-primer/
- [RDFS] Brickley D, Guha, R.V. RDF Schema. 2004; Available from: http://www.w3.org/TR/rdf-schema/
- [OWL] Hitzler P, Krötzsch, M., Parsia, B., Patel-Schneider, P.F., Rudolph, S. OWL 2 Web Ontology Language Primer: W3C; 2009
- [Physio-MIMI]: http://physiomimi.case.edu
- [ASEMR] A. P. Sheth, Agrawal, S., Lathem, J., Oldham, N., Wingate, H., Yadav, P., Gallagher, K., "Active Semantic Electronic Medical Record," in 5th International Semantic Web Conference, Athens, GA, USA, 2006.
- [BioRDF] BioRDF subgroup: Health Care and Life Sciences interest group Available: http://esw.w3.org/topic/HCLSIG_BioRDF_Subgroup
- [TR] A. Ruttenberg, et al., "Advancing translational research with the Semantic Web," BMC Bioinformatics vol. in Press, 2007.

References 2

- [Visage] GQ Zhang et al. VISAGE: A Query Interface for Clinical Research, Proceedings of the 2010 AMIA Clinical Research Informatics Summit, San Francisco, March 12-13, pp. 76-80, 2010
- [Prov] S.S. Sahoo, V. Nguyen, O. Bodenreider, P. Parikh, T. Minning, A.P. Sheth, "A unified framework for managing provenance information in translational research." BMC Bioinformatics 2011, 12:461
- [RO] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C: Relations in biomedical ontologies. Genome Biol 2005, 6(5):R46.
- [BKR] Bodenreider O, Rindflesch, T.C.: Advanced library services: Developing a biomedical knowledge repository to support advanced information management applications. In. Bethesda, Maryland: Lister Hill National Center for Biomedical Communications, National Library of Medicine; 2006.
- T.cruzi project web site: http://wiki.knoesis.org/index.php/Trykipedia

Acknowledgements

- Collaborators:
  - Susan Redline, Remo Mueller, and other members of Physio-MIMI team
  - Rick Tarleton, Todd Manning, Priti Parikh and other members of the *T.cruzi* SPSE team
  - Dr. S. Agrawal and other members at the Athens Heart Center, GA
- NIH Support: UL1-RR024989, UL1-RR024989-05S, NCRR-94681DBS78, NS076965, and 1R01HL087795